

HIGH-THROUGHPUT SEQUENCING AND NATURAL SELECTION: STUDIES  
OF RECENT SWEEP INFERENCES AND A NEW COMPUTATIONAL APPROACH FOR  
TRANSCRIPTION IDENTIFICATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Zhen Wang

August 2014

© 2014 Zhen Wang

ALL RIGHTS RESERVED

HIGH-THROUGHPUT SEQUENCING AND NATURAL SELECTION: STUDIES  
OF RECENT SWEEP INFERENCE AND A NEW COMPUTATIONAL APPROACH FOR  
TRANSCRIPTION IDENTIFICATION

Zhen Wang, Ph. D.

Cornell University 2014

Short-read high-throughput sequencing is the most popular approach to collect massive amount of DNA sequence data at declining cost in nearly all fields of current biological studies. Its many varieties have been employed for different research purposes, e.g. genomic sequencing for variant detection, RNA-seq for transcriptome profiling, etc. However, the individual reads and the resulting called sequences frequently have missing and error-prone base calls, and appropriate corrections and evaluations are necessary for drawing conclusions.

I examined how missing data and sequence errors affect the power and prediction accuracy of two frequently used methods for the inference of recent positive selection from such datasets. I showed that variant-frequency based method, SweepFinder, is very sensitive to data quality and its sensitivity and prediction accuracy are greatly compromised by missing data or sequence errors. In contrast, the haplotype-based method, iHS, is very robust to missing data and sequence errors and is able to efficiently detect signals of recent selective sweeps with very low false discovery rate. I then applied four different computational approaches on the high-throughput resequencing data of a 2.1 Mbp segment of *Drosophila melanogaster* X chromosome to compare and discuss their performances. The study emphasized the relative advantages of linkage disequilibrium-based methods in detecting recent sweeps relative to site frequency-based approaches when applied on incomplete data.

There are also many challenges in other applications of high-throughput sequencing, including discoveries of novel transcription active regions (TARs) in RNA-seq analysis. Here, I present a flexible statistical program, HPIBD (HMM-based Peak Identification and Boundary Definition) for de novo analysis of RNA-seq datasets. It avoids the use of arbitrary read-depth cutoffs and has built-in tolerance to read gaps. It is able to statistically make TARs predictions, estimate peak boundaries and evaluate the confidence in the prediction. I implemented the model and showed that HPIBD has robust performance under various validations and with benchmark to Cufflinks.

## **BIOGRAPHICAL SKETCH**

Zhen Wang was born in Beijing, China. But quickly after that, he started to move around in China with his parents, and ended up stayed in Xi'an, the capital of Shaanxi Province, China, which has over six thousand years in history. After high school, he was admitted to a top science and technology university in China, Tsinghua University, majoring in Biology and minoring in Computer Science. He then came to US and became a graduate student of Genetics, Genomics and Development Program in Department of Molecular Biology and Genetics at Cornell University since 2007. He recently started a full-time job at Bloomberg L.P. focusing on portfolio risk modeling and estimation in September 2013.

## **ACKNOWLEDGMENTS**

I want to thank my PhD advisor Charles (“Chip”) F. Aquadro for his guidance in both academic training and non-academic help throughout my entire PhD, and thank Andrew G. Clark and Alon Keinan for their support and helpful feedback on my projects, manuscripts and this thesis. I also want to thank Carlos D. Bustamante and Richard T. Durrett for their help in my early PhD studies.

# TABLE OF CONTENTS

BIOGRAPHICAL SKETCH .....	I
ACKNOWLEDGMENTS .....	II
TABLE OF CONTENTS .....	III
LIST OF TABLES .....	VII
LIST OF FIGURES .....	VIII
Chapter 1. Introduction .....	1
1.1 Population genetics of adaptation and challenges presented by next-gen sequence datasets .....	1
1.2 Genome sequencing technologies and the challenges of analyzing next-gen sequence datasets .....	7
1.2.1 Modern sequencing technologies .....	7
1.2.2 Errors, missing data and their influence on population genetic investigations .....	9
1.3 RNA sequencing .....	12
Chapter 2. Haplotype-based methods are robust to low-quality high-throughput genome sequencing data for identifying recent selective sweeps .....	15
2.1 Introduction .....	15
2.2 Materials and Methods .....	17
2.2.1 Forward Simulations .....	17
2.2.2 Emulation of Illumina HiSeq 2000 Short reads, Short-read Mapping and SNP Calling .....	19
2.2.3 Emulation of Random Missing Data .....	20
2.2.4 SweepFinder and iHS Calculation and Sweep Evaluation .....	20

2.2.5 ROC Curve Estimation .....	22
2.3 Results .....	22
2.3.1 Power to detect targets of varying strength of selection.....	22
2.3.2 Accuracy in locating targets of selection.....	27
2.3.3 Relationship between False Discovery Rate and power to detect sweeps .....	30
2.3.4 Effect of sample size on low-quality datasets.....	32
2.4 Discussion.....	34
2.5 Acknowledgments.....	39
Chapter 3. Computational Inference of selective sweeps in a 2 Mbp region in <i>Drosophila melanogaster</i> X chromosome.....	40
3.1 Introduction.....	40
3.2 Materials and Methods .....	42
3.2.1 <i>Drosophila melanogaster</i> chromosome X 2 Mb resequencing data .....	42
3.2.2 Computational approaches.....	43
3.2.3 Determining significance of the statistics .....	45
3.3 Results .....	46
3.4 Discussion.....	53
Chapter 4. An HMM-based program for peak detection and estimation of transcript boundaries from high-throughput transcriptome sequencing data.....	57
4.1 Introduction.....	57
4.2 Materials and Methods.....	58
4.2.1 HPIBD framework.....	58
4.2.2 Notations .....	60



4.2.3 Statistically Modeling .....	61
4.2.4 Model parameterization optimization and estimation .....	63
4.2.5 Flexibility .....	64
4.2.6 Implementation .....	64
4.2.7 Benchmarking with Cufflinks.....	65
4.2.8 <i>Drosophila melanogaster</i> Tiling Array data and RNA-seq data.....	65
4.2.9 Differential expression analysis with expression profiling tiling array data....	66
4.2.10 <i>Drosophila melanogaster</i> reference genome, FlyBase annotation, and EST data .....	66
4.2.11 Simulation of RNA-seq data .....	67
4.3 Results .....	71
4.3.1 HPIBD performance against tiling array results.....	71
4.3.2 HPIBD performance against EST data and Fly annotations .....	72
4.3.3 Benchmark of HPIBD with Cufflinks on experimental RNA-seq data.....	72
4.3.4 Benchmark of HPIBD with Cufflinks on simulated RNA-seq data .....	79
4.4 Discussion.....	83
Appendix A: Analysis of DPGP V1 and V2 datasets reveals numerous sequencing biases and errors .....	87
A.1 Introduction.....	87
A.2 Materials and Methods.....	89
A.2.1 Sanger Sequences of <i>Drosophila melanogaster</i> .....	89
A.2.2 DPGP samples .....	89
A.2.3 SweepFinder and iHS Calculation and Sweep Evaluation .....	90

A.3 Results .....	92
A.3.1 Sequencing errors and biases in DPGP V1 compared with Sanger sequences	92
A.3.2 Sequencing errors and biases in DPGP V2 .....	104
A.4 Discussion.....	112

## LIST OF TABLES

Table 3.1 Top putative regions or genes identified to be targets of recent selective sweeps by each method. ....	50
Table 3.2 Results of McDonald-Kreitman test of top putative genes. ....	52
Table 4.1 Simulation evaluations of HPIBD performance.....	81

## LIST OF FIGURES

Figure 2.1: Robustness of SweepFinder (SF) and iHS to missing and erroneous base calls. .....	26
Figure 2.2: Prediction accuracy of SweepFinder and iHS to missing and erroneous base calls.....	29
Figure 2.3: Relationship between iHS power and false discovery rate. ....	31
Figure 2.4: Effect of sample size and missing data on the power of iHS to detect recent selective sweeps. ....	33
Figure 3.1: Top putative targets of recent positive selection inferred by various computational approaches.....	49
Figure 4.1: Graphic representation of HPIBD. ....	69
Figure 4.2: Comparison of Cufflinks and HPIBD on sensitivity of finding TARs. ....	74
Figure 4.3: Distance between predicted and annotated TSS at 5'-UTR. ....	77
Figure 4.4: Distance between predicted and annotated TES at 3'-UTR.....	78
Figure A.1: Example screen shot of comparison of DPGP sequences (above black line) and Sanger sequences (below black line). ....	94
Figure A.2: Spatial pattern of missing data distribution of a 2 Mb window on X chromosome. .	95
Figure A.3: Polarized site-frequency spectrum (SFS) of 22kb on chromosome X comparing Sanger sequences (red), raw Solexa sequences (blue) and Solexa sequences with $Q \geq 20$ filtering (black). ....	98
Figure A.4: DPGP V1.0 sequences contain substantial transversion errors and fewer transition errors. ....	99
Figure A.5: DPGP V1.0 has no bias toward sequencing a particular type of nucleotide acid with low quality. ....	100
Figure A.6: Performance of SweepFinder on datasets with varying level of quality and missing data.....	102

Figure A.7: Identification of putative targets of recent selective sweeps by iHS using DPGP V1 datasets with varying quality filtering .....	103
Figure A.8: Number of <i>D.melanogaster</i> individuals sequenced in each population in DPGP V2.....	105
Figure A.9: Base call qualities of DPGP V2 sequences.....	106
Figure A.10: Analysis of a 2 Mb region in RG population with all 27 individuals. ....	107
Figure A.11: Proportion of variant sites that can be used under varying quality filtering and sample size requirements.....	110
Figure A.12: Putative targets identified by iHS for the X chromosome in Rwanda population.....	111

## Chapter 1. Introduction

### **1.1 POPULATION GENETICS OF ADAPTATION AND CHALLENGES PRESENTED BY NEXT-GEN SEQUENCE DATASETS**

Population geneticists have long been exploring how natural selection shapes genetic variation. Most of the methods used to detect positive selection can be classified into two categories: divergence-based methods and polymorphism-based methods. A classic example of divergence-based methods is to compare the ratio of nonsynonymous mutations per nonsynonymous site to synonymous mutations per synonymous site ( $d_N/d_S$ ) to detect recurrent selective fixations. Synonymous mutations are assumed to be nearly neutral. So  $d_N/d_S = 1$  is expected for a neutral locus, and  $d_N/d_S > 1$  for a positively selected region (YANG 1997; HUELSENBECK and RONQUIST 2001; YANG 2007). There have been modifications to this method as well, such as comparing the noncoding mutations with  $d_S$  from coding regions, or testing for evidence of lineage-specific acceleration of divergence (CLARK *et al.* 2003; WONG and NIELSEN 2004; HOLLOWAY *et al.* 2008). In general, these approaches are powerful if recurrent selection has occurred at multiple sites along a species lineage.

Alternatively, polymorphism-based methods have been developed to detect recent selective events. A signature of recent selection is the so-called “selective sweep” in which genetic variation at nearby neutral loci is lost or its frequency spectrum skewed

as an advantageous allele goes to fixation. The magnitude of the effect of a selective sweep decreases with genetic distance from the target of selection. Expected signatures of a selective sweep (e.g. an excess of rare alleles and high frequency derived alleles) are widely used to help identify targets of positive selection.

One popular way to detect positive selection using polymorphism data is the empirical genomic scan approach. In brief, a large dataset of loci in the genome are collected and summarized into one (or more) summary statistics (e.g. nucleotide diversity,  $\pi$ ), which are used to construct empirical distribution(s). Outliers of the distributions of these summary statistics, or outliers that could not be fit into a plausible demographic model, are identified as candidate selected targets (GLINKA *et al.* 2003; DUMONT and AQUADRO 2005; OMETTO *et al.* 2005; WRIGHT *et al.* 2005; THORNTON and ANDOLFATTO 2006). However, the reliability of this empirical method is affected by many factors. For example, false positive and false negative rates can be unacceptably high if the proportion of true selected sites differs much from the arbitrary cutoff of significance, or if selection has acted on standing variation that was neutral in the population before it was selected (TESHIMA *et al.* 2006).

Model-based approaches are also widely used to identify recently selected loci. They typically test the selection model against the non-selective model and accept the one with significantly greater likelihood. Kim and Stephan (2002) developed a composite likelihood approach that can detect and locate positive selection using site-

frequency spectrum (SFS) information. Spatial variation in SFS is considered and a likelihood ratio test is then performed between a simple sweep model and the standard neutral model. Nielsen *et al.* (2005b) developed a composite likelihood to identify potential regions under positive selection using genome-wide genotyping data, based on the composite likelihood ratio test (CLRT) of Kim and Stephan (2002), that has been used in many population genetic surveys, e.g. (CARLSON *et al.* 2005; WILLIAMSON *et al.* 2007; NIELSEN *et al.* 2009). Nielsen *et al.*'s method utilizes putatively neutral regions in the genome as “background” in substitution for a specific demography model. SweeD is another modification of the Nielsen *et al.*'s method that adopts a variable-size window and incorporates invariant sites to reduce the sensitivity to regions of low SNP density (PAVLIDIS *et al.* 2013). However, the performance of both likelihood methods decreases dramatically and predictions of the regions harboring the selective target become very large with incomplete sequence data (JENSEN *et al.* 2008b).

One main problem with such polymorphism-based methods is the fact that relatively high density of SNPs is needed to make inference and power is dramatically compromised in detecting partial sweeps. Also the predicted patterns of genetic variation during and after selection can be confounded by many external factors such as non-equilibrium demography (JENSEN *et al.* 2005). Variation in recombination and mutation can also change nucleotide diversity regionally and mimic the predicted “selective signature”. There have been efforts to adjust for demography using



information from neutral loci (AKEY *et al.* 2002; BUSTAMANTE *et al.* 2002; LI and STEPHAN 2005; NIELSEN *et al.* 2005a), based on the prediction that selection would only affect a small proportion of the genome that is genetically linked to selective sites (HUDSON *et al.* 1987; GALTIER *et al.* 2000). Previously efforts were made to estimate demographic history based on a large collection of loci, and then the more realistic non-selective model is tested against selection (ANDOLFATTO and PRZEWORSKI 2000; WALL *et al.* 2002; NIELSEN *et al.* 2005a; WRIGHT *et al.* 2005). Though this method provides us a lot of insights into how genomes evolve, demographic parameters could be inaccurate sometimes either because too few loci are included in the estimation or because of inappropriate assumptions.

In contrast to SFS-based methods, haplotype-based methods such as iHS, XP-EHH and Omega, examine haplotype structure across the genome (SABETI *et al.* 2002; KIM and NIELSEN 2004; VOIGHT *et al.* 2006; JENSEN *et al.* 2007b; SABETI *et al.* 2007; PAVLIDIS *et al.* 2010). These methods estimate the age of a core haplotype by its association with nearby alleles. Core haplotypes with high iHS values and high frequencies are indicators of a past mutation that was driven to high frequency or fixation faster than neutral expectation. Thus, such loci might be targets of recent selective events. iHS is most used to detect partial sweeps within a population though its power on complete sweeps is unclear, while XP-EHH has more power to identify complete sweeps with respect to two populations. Such LD-based approaches have built-in correction for variation in

recombination and are believed to be relatively robust to demography. Omega utilizes the high correlation between SNPs within each side of the complete sweep, but not between, to identify complete sweeps. However, the accuracy of haplotype-based approaches to locate physical positions of recent sweeps are limited by both SNP density and how fast LD decays, and typically have much wider prediction ranges with complete data compared with SFS-based methods (GROSSMAN *et al.* 2010).

Previously, exploring the evolutionary impact on natural variability at a genomic level was severely limited because the cost and labor needed to sequence entire genomes were insurmountable for a single lab. An alternative was to sequence regions for a sample of loci, assuming this was sufficient to perform tests and draw conclusions. However, this subset of loci and variants may not be good representative of the genome, and thus bias the results (MARIONI *et al.* 2008). Next-generation sequencing, which enables the fast and relatively low cost of sequencing genomes, is making nearly unbiased genome-wide sampling possible. These methods will allow the collection of genetic variation at high resolution for both coding and noncoding regions, which could lead to a better understanding of regulatory sequence evolution. There have been whole-genome studies published in *Saccharomyces cerevisiae* (DONIGER *et al.* 2008), *Arabidopsis thaliana* (OSSOWSKI *et al.* 2008), *Caenorhabditis elegans* (HILLIER *et al.* 2009), *Drosophila melanogaster* (DAINES *et al.* 2009; LANGLEY *et al.* 2012; MACKAY *et al.* 2012; POOL *et al.* 2012), and *Homo sapiens* (WANG *et al.* 2008; WHEELER *et al.* 2008; AHN *et al.* 2009;

GENOMES PROJECT *et al.* 2010; GENOMES PROJECT *et al.* 2012). However, due to the error-prone nature of the high-throughput, short-read sequencing technologies, a lot of the available datasets have extensive missing data and sequence-base call uncertainty. The large size of the datasets (typically genome-wide) requires computationally feasible methods to do population genetics studies. But the low data quality will impose constraints on the methods that can be used, particularly those that require full site-frequency spectra and unbiased ascertainment of low frequency variants.

Due to the rapid progress in next-generation sequencing technologies the shift in the field has been toward the identification of the targets of natural selection on a genomic scale. While the cost of high-throughput sequencing continues to decline, population genetic data at genomic scale are expanding into both model and non-model organisms. Several polymorphism-based approaches were developed recently to try to make quantitative inferences about selection such as selective strength and the rate of recurrent sweeps using *Drosophila* data (LI and STEPHAN 2006; ANDOLFATTO 2007; STEPHAN and LI 2007; JENSEN *et al.* 2008a; PAVLIDIS *et al.* 2010). However, these studies have yielded some strongly conflicting estimates (e.g. the average genomic selection coefficients of adaptive mutations in *Drosophila melanogaster* range from very weak, e.g.  $s=0.00001$ , to very strong, e.g.  $s=0.01$ ). With more complete and more unbiased data generated by next-generation sequencing, future population genetic surveys will be able

to detect weaker sweeps from denser data, thus help distinguish between different selective scenarios.

## **1.2 GENOME SEQUENCING TECHNOLOGIES AND THE CHALLENGES OF ANALYZING NEXT-GEN SEQUENCE DATASETS**

### **1.2.1 Modern sequencing technologies**

Modern capillary-based Sanger sequencing has always been of the highest accuracy since its invention in the early 1990s (SANGER 1988; SWERDLOW and GESTELAND 1990; SWERDLOW *et al.* 1990; HUTCHISON 2007). This semi-automated sequencing technique can have limited level of parallelization of simultaneously running 96 or 384 capillaries and was applied in shotgun de novo sequencing of random DNA fragments in addition to the tradition of PCR amplification of a target region (SHENDURE *et al.* 2004). By using fluorescently labeled reversible terminators (ddNTPs), Sanger sequencing typically achieves about 1,000 bp in read length and has very low error rate of  $10^{-5}$  to  $10^{-6}$  per base (HUNKAPILLER *et al.* 1991; EWING and GREEN 1998; EWING *et al.* 1998). Despite its much higher cost than more modern high throughput platforms, Sanger sequencing is still widely used in current Biological analysis that requires very high quality of data, e.g. verification of potential genetic variations.

The short-read large-scale sequencing is the most popular approach to collect massive amount of data at declining cost in nearly all fields of current biological studies

(see (ZHANG *et al.* 2011) for methodology review). While new technologies are being developed and tested (e.g. single-molecule sequencing), Illumina sequencing (BENTLEY *et al.* 2008), 454 Life Sciences (Roche) pyrosequencing (MARGULIES *et al.* 2005), Applied Biosystems SOLiD sequencing (MARDIS 2008) are the dominant examples of sequencing platforms available in current commercial market. The potential applications of next-generation (NGS, also called second-generation) sequencing techniques have been drastically strengthened by more accurate whole genome assemblies in model organisms and tremendous improvement for non-model organisms (DONIGER *et al.* 2008; OSSOWSKI *et al.* 2008; WANG *et al.* 2008; WHEELER *et al.* 2008; AHN *et al.* 2009; HILLIER *et al.* 2009; GENOMES PROJECT *et al.* 2010; SCHWARTZ *et al.* 2010; GENOMES PROJECT *et al.* 2012; BRADNAM *et al.* 2013). Such assemblies, which are still undergoing rapid improvement, serve as a reference for short reads mapping, and various downstream bioinformatic analysis, including genetic variation discovery, RNA expression analysis, DNA-protein interaction and epigenetic surveys (KOBOLDT *et al.* 2013; RIVERA and REN 2013).

One major application of NGS is variant detection. The advantage of variant identification by sequencing is that most variants, common or rare, known or novel, nucleotide or structural, can be discovered with corresponding sequencing methodology, sequencing depth and coverage and appropriate bioinformatic software. As previously mentioned, a reliable reference genome often serves as a starting point, so that various variant calling algorithms can be employed for downstream applications

such as population genetic surveys (e.g. SNP calling (LI *et al.* 2008a; LI *et al.* 2008b; KOBOLDT *et al.* 2009; LANGMEAD *et al.* 2009; LI *et al.* 2009b; MCKENNA *et al.* 2010; SHEN *et al.* 2010; DEPRISTO *et al.* 2011; LIU *et al.* 2012); indel detection (YE *et al.* 2009; EMDE *et al.* 2012; ONMUS-LEONE *et al.* 2013)), etc.).

### **1.2.2 Errors, missing data and their influence on population genetic investigations**

Due to the complex nature in the sequencing chemistry and subsequent analysis, errors and missing data could stem from any of the steps. There are three major sources of sequencing inaccuracy: sequencing errors, bioinformatic errors and missing data. Though typically sequencing accuracy can be improved by incorporating more individuals (larger sample size) (GENOMES PROJECT *et al.* 2010; GENOMES PROJECT *et al.* 2012; LANGLEY *et al.* 2012; MACKAY *et al.* 2012; POOL *et al.* 2012), extending the coverage of the genome (deep sequencing) (BENTLEY *et al.* 2008), and applying advanced bioinformatic algorithms (e.g. analyze all reads from all samples together instead of individual-specific base calling) (STONE 2012), there is always trade-off between quantity and quality, especially for organisms with large genome size. Under given budget and specific research goals, sequencing coverage and sample size are always evaluated against the statistical power and errors that will be used for lower-coverage datasets, and such trade-off should be taken into account in the experimental design.

The stage where error occurs will determine the observed error frequency in the sequencing results. For example, failure of PCR amplification before sequencing and asynchronous strand elongation/synthesis during actual sequencing would lead to missing data and potential sequencing errors, respectively. Furthermore, though generally it was believed that DNA fragments in the prepared sample libraries are sequenced randomly, there is much evidence showing that fragments of either very high or very low GC-content tend to be under-represented (DOHM *et al.* 2008; OSSOWSKI *et al.* 2008). Also, both random, single-read erroneous base calls and nonrandom errors across multiple reads are observed (KEIGHTLEY *et al.* 2009). Both biases will skew the site frequency spectrum (SFS) towards rare alleles, which leads to an excess of rare variants (JOHNSON and SLATKIN 2008). Such biased SFS will greatly compromise the performance of many population genetic computational approaches of inferring adaptive selection that are based on it. However, applying more stringent quality filtering is not an optimal solution (JOHNSON and SLATKIN 2006; JOHNSON and SLATKIN 2008), because it will also bias SFS as well by excluding true SNPs from the analysis.

Mapping errors are another source of erroneous base calls and variant inference. The length of the reads from next-generation sequencers is much shorter than traditional Sanger sequences, e.g. currently up to 150bp on Illumina platforms. It is very challenging to map such short reads to the reference genome accurately with variants, especially in highly repetitive regions (LI *et al.* 2008a; LI *et al.* 2008b; LANGMEAD *et al.*

2009; LI *et al.* 2009b). Current mapping algorithms discard reads with mismatches of more than 2 or some user specified number (LI *et al.* 2008a). As a result, this will tend to discard reads with alternative alleles and thus lead to: (1). Less coverage and depth for alternative alleles; (2). Potentially inaccurate variant calling for alternative alleles; (3). Bias in sample SFS towards those alleles found in the reference, which are typically the high frequency alleles since the reference consensus is often obtained from multiple individuals. The first two can also underestimate the region nucleotide diversity level and may compromise the performance of some computational methods due to dependence on SNP density. By employing pair-end sequencing techniques that enhances the mapping capabilities and incorporating known polymorphisms into the reference sequence, such reference sequence bias can be alleviated. Pair-end sequencing can also be particularly useful to determine the phasing status in population genetic surveys when individuals are not inbred.

Missing data can be the result of sequencing errors and later quality filtering. It is also a concern that not both alleles from an individual are sequenced, thus leading to missing data at one allele, especially in low-coverage datasets (BENTLEY *et al.* 2008). Missing data will cause variation in sample size at different sites along chromosomes and introduce possible biases in population parameter estimation (e.g. nucleotide diversity, etc.) and hurt the power of identifying the footprints of selection. While imputing the missing SNPs is possible (LYNCH 2009; MARCHINI and HOWIE 2010; PORCU



*et al.* 2013), it will also introduce biases in population genetic analysis such as bias the SFS against singletons since they cannot be imputed (BHANGALE *et al.* 2008).

Many potential errors and biases in next-generation sequencing techniques have been reported, but it is still unclear how population genetic analysis and current computational tools will be affected if applied on such datasets. Finding the optimal experimental design for a given research goal is a particular important question to keep in mind. In Chapter 2, we performed analysis and comparisons how different computational approaches would behave with potential challenges from the next-generation datasets and in Chapter 3 we compared their predictions by applying them to a *Drosophila melanogaster* dataset to infer recent selective sweeps.

### **1.3 RNA SEQUENCING**

RNA-seq is one of the many applications as to how next-generation sequencing can be employed to obtain useful "big data" in biological studies other than genome re-sequencing. Compared to microarrays, RNA-seq can have a resolution as fine as 1 bp and does not need prior information to detect novel transcripts. As it is becoming more and more cost effective with rapid advancements in high throughput technologies, its merits such as requiring less starting materials, low background noise, high sensitivity and capability to retrieve expression information without a reference genome is making

its way to most transcriptome projects (WANG *et al.* 2009; MARGUERAT and BAHLER 2010; OZSOLAK and MILOS 2011).

There are many novel applications of RNA-seq in research, such as genome-wide expression profiling (AGARWAL *et al.* 2010), differential expression (TRAPNELL *et al.* 2010), mapping transcription starting and ending sites (NI *et al.* 2010), strand-specific expression profiling (LEVIN *et al.* 2010), etc.

Despite the tremendous benefits it brings to transcriptome studies, it has been known that NGS libraries for various applications can contain biases and errors that would potentially harm the quality of NGS datasets and result in challenges from initial sequencing to downstream bioinformatic analysis and interpretation (WANG *et al.* 2009). Such biases and errors will likely cause low coverage for some fragments (for example, GC-content bias for fragments with extreme GC contents), leading to incomplete coverage and missing data in the transcriptome, thus impose greater challenges for following analysis. Another important bias in RNA-seq is the heterogeneity in sequencing alternative alleles and nucleotides across an expressed region. The former will lead to uneven number of reads for different alleles, thus low or no coverage of some alleles. The latter will result in varying sequencing depth at different nucleotide sites along a transcript. Such heterogeneities in coverage may strongly affect the reliability and accuracy of expression estimates and other inferences for transcripts (BULLARD *et al.* 2010; HANSEN *et al.* 2010; LI *et al.* 2010a).

Further, intermediate to lowly expressed genes are always challenges for reliable detection and accurate quantification thus leading to inaccurate inferences and false discoveries (AUER and DOERGE 2010). Deeper sequencing coverage helps improve the sensitivity and reduce the large variance, but effects are limited. For instance, the low-expression *doublesex* gene, which functions in sexual dimorphism in flies, was not identified by deep coverage in RNA-seq of the modENCODE embryonic samples while it is known to be expressed at this stage (GRAVELEY *et al.* 2011).

Though RNA-seq has been widely used in many large-scale projects, bioinformatic tools are relatively limited and usually have constraints. For example, almost all analytic programs involve arbitrary read depth cutoffs as one of the steps to label regions as containing mapped reads. This approach can be confounding when read mapping criteria are relaxed and more reads are mapped incorrectly causing higher level of background noise. Also, even for the very few model organisms that have relatively well-annotated genomes, their annotations are often still incomplete and reliable methods to identify exon-exon splicing junctions are needed (GUTTMAN *et al.* 2010; TRAPNELL *et al.* 2010). Knowing that the RNA-seq data contain background noise and incomplete information, the particular bioinformatic tools used to analyze the datasets can differ dramatically in their predictions and researchers need to be careful to choose the appropriate software for their scientific purposes.

## **Chapter 2. Haplotype-based methods are robust to low-quality high-throughput genome sequencing data for identifying recent selective sweeps**

### **2.1 INTRODUCTION**

Next-generation sequencing enables the fast and relatively low cost of re-sequencing genomes in both model (DAINES *et al.* 2009; XIA *et al.* 2009; GENOMES PROJECT *et al.* 2012; HELYAR *et al.* 2012; LANGLEY *et al.* 2012; MACKAY *et al.* 2012; POOL *et al.* 2012) and non-model species (STAPLEY *et al.* 2010; ERSOZ *et al.* 2012; HELYAR *et al.* 2012). While large sample size is expected to improve the statistical power of subsequent population genetic surveys, next-generation sequencing datasets typically have greater uncertainty in base and variant calling, which often leads to low-quality datasets with missing data (HELLMANN *et al.* 2008; JOHNSON and SLATKIN 2008; LYNCH 2008; LYNCH 2009; AIRD *et al.* 2011; GUO *et al.* 2012; TESCHENDORFF *et al.* 2013). Some researchers have tried to circumvent some of these problems by restricting their site-specific analysis to only individuals with sufficient read coverage (JIANG *et al.* 2009) although this approach usually reduces sample size and further increases missing data. Low-pass experimental design, stringent SNP calling or lack of good reference genomes, such as in most non-model organisms, can also lead to extensive missing data.

Population genetic approaches for identifying evidence for positive selection in genome scans were initially developed for small-scale high-quality datasets. Jensen *et al* (JENSEN *et al.* 2008a; JENSEN *et al.* 2008b) showed that the site-frequency spectrum (SFS) based method, Composite Likelihood Ratio Test (CLRT), had a dramatic reduction in both power to detect recent positive selection and precision in locating the target of selection when applied to segmentally sequenced regions. It is thus a concern that many of the methods will not perform efficiently and accurately on next-gen datasets with significant extent of missing data from scattered nucleotide sites as well.

Compared to SFS-based methods, the integrated Haplotype Score (iHS) approach takes advantage of linkage disequilibrium information and examines haplotype structure to detect recent selective sweeps (VOIGHT *et al.* 2006; SABETI *et al.* 2007; GROSSMAN *et al.* 2010). Since linkage disequilibrium (LD) and haplotype structure depend mostly on common SNPs, and sequence errors tend to be rare putative variants (JOHNSON and SLATKIN 2008; POOL *et al.* 2010), we reasoned that iHS should be robust to missing data and sequence errors in next-gen population datasets.

The purpose of our study is to determine which of these two polymorphism-based approaches are more appropriate for low-quality datasets to infer targets of recent positive selective events. Using simulated datasets, we investigated both complete and partial selective sweeps and estimated how each method's performance was affected by missing data and sequence errors under different combinations of selected allele

frequency, selective strength and sample size. Missing data and sequence errors in the datasets dramatically reduced SweepFinder power and target prediction. In contrast, iHS was robust to missing data and sequence errors under all selective strength scenarios simulated. We also show that signals of selection can be efficiently detected by iHS with relatively small sample size of 24 chromosomes for a species such as *Drosophila melanogaster*. While increasing sample size helps improve power to identify selection, missing data in the sequences offsets these benefits

## **2.2 MATERIALS AND METHODS**

### **2.2.1 Forward Simulations**

We used the forward simulator SFS\_CODE (HERNANDEZ 2008) in all simulation studies in this paper. We first simulated over ten thousand population replicates under the Wright-Fisher Model with constant population size, random mating and no natural selection. The estimated effective population size of  $1 \times 10^6$  for *Drosophila melanogaster* was used in the model (KREITMAN 1983). A mutation rate of  $1 \times 10^{-8}$  per base pair per generation was used and 10 kb segments were simulated given the fact that SNP are of much higher density and LD decays very fast in *D. melanogaster* (MACKAY *et al.* 2012). The 10 kb simulated segments for *D. melanogaster* on average have more than 420 SNP sites even under strong selective sweeps with  $2N_s=1000$  and about 500 SNP sites with  $2N_s=500$  given the theta value and effective population size simulated. Analysis of our

simulated datasets also confirms that they show a decay of LD typically within 1 kb under neutrality and within 2 kb with strong selection of  $2N_s=1000$  (data not shown). Recombination rates on the 10 kb segment varied every 2 kb, and were sampled randomly with replacement from the pool of estimated recombination rates from *Drosophila melanogaster* chromosome X (FISTON-LAVIER *et al.* 2010). We simulated the evolutionary trajectory of the population and randomly drew sample sizes of 50, 37 and 24 chromosomes.

Varying strengths of selective sweeps from strong to weak (selective advantage  $s=[0.0005, 0.00025, 0.0001]$ , equivalent to  $2N_es=[1000, 500, 200]$ ) were simulated under the same set of population parameters described above for varying sample size of 50, 37 and 24 chromosomes. For complete sweeps, we restricted the time since its completion to sampling to be within  $0.005N_e$  generations. The population allele frequency was determined for the selected allele and sweeps with similar allele frequencies were grouped for downstream analysis.

For subsequent analysis, we first edited the simulated sequences so that invariant sites and the ancestral state for each polymorphic site matched the reference sequence of *Drosophila melanogaster* X chromosome downloaded from FlyBase release v5.3 (MARYGOLD *et al.* 2013). These edited simulated sequences were further modified to emulate either Illumina HiSeq 2000-specific missing data and sequence errors, or random missing data, as detailed in the next two sections.

### 2.2.2 Emulation of Illumina HiSeq 2000 Short reads, Short-read Mapping and SNP Calling

To emulate Illumina-specific missing data and sequence errors, we used SimSeq program (EARL *et al.* 2011) to fragment the sample sequences and generate short reads according to Illumina HiSeq 2000 error profile, followed by BWA reassembly to the *Drosophila melanogaster* chromosome X. We assumed homozygous individuals in short-read generation and used the short-read simulation program to emulate computational fragmentation and generate short reads that would be obtained from Illumina HiSeq 2000 platform. SimSeq takes advantage of human-derived HiSeq 2000 reads to train its error model and then use the error profile to emulate the HiSeq 2000 platform by introducing sequencing errors computationally. Paired-end reads of 100 bp were sampled from average insert size of 500 bp (standard deviation of 50 bp) for each homozygous individual at 10X coverage and output in BAM alignment file format. Duplicate probability was set to be 0.01 and indels were not modeled. The BAM files were then converted to SAM format using SAMtools (LI *et al.* 2009a) and subsequently translated into raw reads of FASTQ format using the SAMtoFastQ command line tool in the Picard Tools (v1.56) package (<http://picard.sourceforge.net/>).

Short-read libraries were then aligned to the reference sequence using BWA (LI and DURBIN 2009) with default parameters. And SNPs were called using SAMtools (LI *et al.* 2009a). The varying number of segregating sites was controlled by adjusting the



quality cutoff in SNP calling. With an average number of 400 segregating sites representing 100% of the true segregating sites present, we found that 60% segregating sites roughly matched quality filtering with Phred  $Q>20$  in SNP density (as observed in the Drosophila Population Genomics Project Illumina dataset (release 1.0) for *D. melanogaster*; <http://www.dpgp.org/>).

### **2.2.3 Emulation of Random Missing Data**

An alternative approach to simulate missing data is to use a simple random model. Here, polymorphic sites in simulated population samples were randomly masked with equal probability with in-house scripts without generation of short-read libraries.

### **2.2.4 SweepFinder and iHS Calculation and Sweep Evaluation**

We used SweepFinder (NIELSEN *et al.* 2005b) to perform the Composite Likelihood Ratio Test (CLRT). SweepFinder takes the polymorphism data and compares the likelihood of the model with a sweep to the likelihood of the neutral model without selection but based on genome background allele frequencies. A grid size of 10, equivalent to 1 kb, was used in the analysis. The critical value was calculated from the empirical distribution consisting of likelihood ratios from the ten thousand neutral simulations generated above given FDR=1%. Statistical significance was then evaluated

and a sweep was “detected” if the likelihood ratio of the sample was greater than the critical value determined as described above.

To confirm the simulated samples under selection contained sufficient SNPs for SweepFinder to construct background allele frequencies, we performed two different sweep detection and significance evaluation procedures after matching SNP density of simulated samples to Drosophila Population Genomics Project (DPGP) Illumina dataset (release 1.0 for *D. melanogaster*; <http://www.dpgp.org/>). In the first approach, we applied SweepFinder on sweep simulated and edited samples, and constructed background site frequency spectrum (SFS) directly from neutral SNPs in the sample sequences. Alternatively, we used neutrally simulated samples with all of the same parameters (except that  $s=0$ ) to construct the background allele frequencies for SweepFinder. We obtained identical inferences of selection using both sets of neutral reference SNPs.

The iHS program was downloaded from Pritchard lab site (<http://hgdp.uchicago.edu/>). iHS utilizes haplotype structure to examine the existence of extended haplotypes on the genetic map (VOIGHT *et al.* 2006). The 99% confidence interval was estimated from the empirical distribution of the ten thousand iHS values calculated from the neutral simulations generated above given FDR=1%. SNPs under a sweep model were deemed to be significant if its iHS value fell outside of the 99% confidence interval. A putative target of selective sweep was considered statistically significant if more than 30% of the SNPs in a 10-SNP sliding window (with step size of 5

SNPs) had significant iHS scores. Power was estimated as the proportion of identified sweeps to the total number of simulated sweeps (calculated from a minimum of 100 simulations). Predicted target sites of selection were defined to be the midpoint of the significant 10-SNP windows (Voight et al. 2006; Grossman et al. 2010).

### **2.2.5 ROC Curve Estimation**

Critical values of SweepFinder and iHS were calculated from empirical distributions under Wright-Fisher model for different FDR values. For selective strength, the corresponding power for each sweep end-point allele frequency was then assessed under each FDR value. The overall power given the FDR was then calculated as the average of the power across all sweep frequencies from 0.1 to 1.0.

## **2.3 RESULTS**

### **2.3.1 Power to detect targets of varying strength of selection**

We first estimated the power of SweepFinder and iHS to identify targets of selection using our simulated population sequence datasets like that obtained from Illumina HiSeq 2000 runs (Fig 2.1 A-C). When selective strength varied from weak to strong ( $s=[0.0005, 0.00025, 0.0001]$ , or equivalently  $[2N_{e}s=200, 2N_{e}s=500, 2N_{e}s=1000]$ ), we found that SweepFinder had high sensitivity (true positive rate) when selection was strong ( $2N_{e}s=1000$ ) and much lower sensitivity when it was weak ( $2N_{e}s=200$ ). As

expected, SweepFinder had almost no power in identifying partial selective sweeps (e.g., with selected allele frequency under 0.8 in the population), but did have relatively high power to detect fixed (complete) sweeps. This is consistent with the fact that this method was developed to detect recently completed sweeps. In all selective scenarios simulated, increasing the extent of missing data and sequence errors had a greater effect on SweepFinder power reduction. Specifically, SweepFinder power decreased from 82.2% to 58.3%, 27.8% and 14.6% under strong selection ( $2N_{es} = 1000$ ) when 0%, 20%, 60% and 75% of the total segregating sites were missing, respectively (Fig. 2.1A). Under weak selection ( $2N_{es} = 200$ ), SweepFinder had power of less than 10% when over 60% of the segregating sites were missing (Fig. 2.1C).

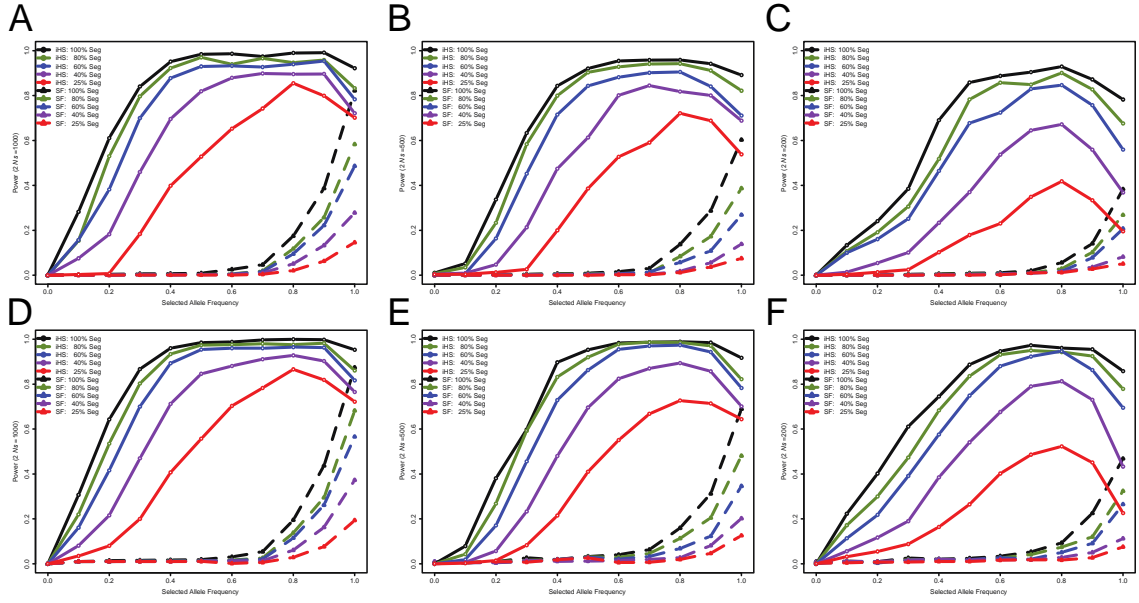
iHS exhibited highest sensitivity to mid- to high frequency sweeps: it began to gain power with sweep alleles with frequency  $>0.1$  and reached a plateau with sweep allele frequency  $>0.5$  before it started to decline slightly with sweep allele frequency  $>0.8$ . iHS was able to reach a very high power of  $>95\%$  for high frequency partial sweeps under medium and strong selections ( $2N_{es}=500$  and  $2N_{es}=1000$ ) and  $>85\%$  under weak selection ( $2N_{es} = 200$ ) using complete data.

Curves with 100%, 80% and 60% segregating sites were quite comparable and plateaued around 80% in power when selection was medium to strong, and around 70% when selection was weak. Surprisingly, under medium to strong selection, iHS still retained power to effectively identify over 60% partial sweeps of mid- to high frequency

when only 25% of total segregating sites were available. With weak selection, further reductions of variant coverage from 60% to 40% and then 25% segregating sites showed remarkable decrease in iHS power from 55.9% to 36.8% and 19.5%, respectively (Fig. 2.1C). Nonetheless, iHS consistently outperformed SweepFinder and remained very robust to missing data and sequence errors. Interestingly, iHS demonstrated more power with missing data on complete sweeps than SweepFinder did, though the former method was initially developed to detect partial sweeps. This retained power appears due to the fact that most of the complete sweeps simulated had fixed very recently ( $<0.005N_e$  generations) and the haplotype structures generated by the sweeps and detected by iHS were mostly retained.

We also tested SweepFinder and iHS sensitivity for these same combinations of missing data and strength selection, but with the missing data generated by random deletion of the SNP sites (referred as random missing hereafter) instead of deletion following the HiSeq 2000 error model (Fig 2.1 D-F). We found that both SweepFinder and iHS had consistent differences in performance with both the random missing and HiSeq 2000 error models, and there was only a slight decrease in power in the random vs the HiSeq 2000 error models (SweepFinder power decreased by 7.4% on average, only considering fixed sweeps, and iHS power decreased by 4.6% on average). We think the minor difference between the two error simulation schemes is because the HiSeq 2000 error model tends to result in greater bias for rare variant frequencies, while randomly

missing data biases the entire SFS. Random missing data thus had a smaller negative performance impact on iHS than on SweepFinder, consistent with the iHS method being overall more robust to missing data.



**Figure 2.1: Robustness of SweepFinder (SF) and iHS to missing and erroneous base calls.**

Sample size of 50 chromosomes were simulated. Power of SweepFinder (SF) (dashed lines) and iHS (solid lines) was estimated as described in Methods. (A)-(C) Error model for short read generation mimicking Illumina HiSeq 2000 platform with  $2N_e s = 1000$  (A),  $2N_e s = 500$  (B), and  $2N_e s = 200$  (C). (D)-(F) To simulate missing base calls, segregating sites (referred as Seg in all legends) were randomly eliminated from the results of complete sequence simulations that used  $2N_e s = 1000$  (D),  $2N_e s = 500$  (E), and  $2N_e s = 200$  (F).

### 2.3.2 Accuracy in locating targets of selection

We also evaluated ability of both SweepFinder and iHS to accurately predict the location (e.g. target) of the selective sweeps with sequencing errors and missing data. Only samples where putative sweeps were predicted were included in the analysis. The distance between the predicted selected site and the true sweep site was used to measure prediction accuracy.

Under strong, medium and weak selection, 80.5%, 74.9%, 69.0% of putative target sites predicted by SweepFinder using complete data fell within  $\pm 1$  kb from the true sweep targets with average distance of 0.69 kb, 0.82 kb and 0.93 kb, respectively (Fig. 2.2A). However, with 80% segregating sites used for analysis, predictions within  $\pm 1$  kb from sweep target dramatically dropped down to 55.9%, 51.2% and 45.2% with average distance of 1.15 kb, 1.23 kb and 1.35 kb, respectively, under strong to medium and weak selection (Fig. 2.2B). On average, 23.2% putative target sites were  $>2$  kb from true targets compared to less than 3% when no missing data were present. Overall, SweepFinder showed improved prediction accuracy under stronger selection, but was greatly compromised when there are missing data.

In contrast, iHS exhibited poorer prediction accuracy with complete data, but accuracy was only slightly further reduced by missing data. The proportion of putative sites predicted by iHS using complete data (or 80% segregating sites) that fell within  $\pm 1$  kb from the true sweep sites were 30.2% (29.4%), 35.6% (34.0%) and 39.5% (38.3%) with



average distance of 1.82 kb (1.90 kb), 1.62 kb (1.67 kb) and 1.51 kb (1.53 kb) under strong to medium and weak selection, respectively. On average, 36.9% of predictions (35.5% for 80% segregating sites) fell more than 2 kb from the true targets when there were no missing data (Fig. 2.2C, D). Interestingly, iHS prediction accuracy actually deteriorated under stronger selection. This result appears to be because stronger selection creates longer haplotypes and thus compromises iHS target prediction accuracy.

For samples where both SweepFinder and iHS predicted putative sweeps, iHS on average lies 0.9 kb more distant with complete data, and 0.4 kb using 80% segregating sites compared to SweepFinder. Though iHS was less accurate than SweepFinder in locating sweep sites under the two scenarios examined, iHS was able to predict sweep sites more accurately than SweepFinder when pronounced proportion of data were missing due to the fact that missing data had much greater negative impact on SweepFinder prediction accuracy than on iHS (data not shown).

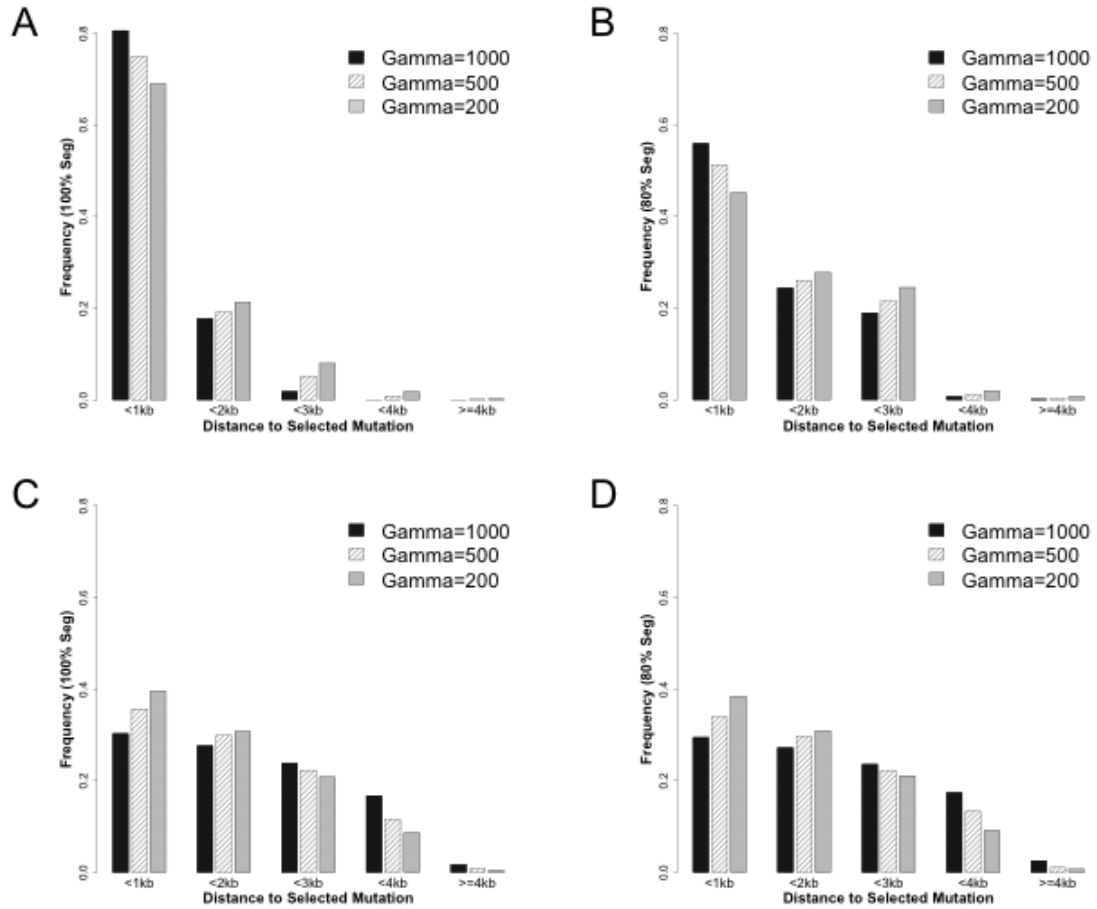


Figure 2.2: Prediction accuracy of SweepFinder and iHS to missing and erroneous base calls.

Accuracy in locating selective sweeps was measured by distance between the inferred sites and true sweep sites. Selective strength ( $2N_e s$ ) varies from strong (black,  $2N_e s = 1000$ ) to medium (blue,  $2N_e s = 500$ ), to weak (red,  $2N_e s = 200$ ). (A) SweepFinder with 100% segregating sites; (B) SweepFinder with 80% segregating sites; (C) iHS with 100% segregating sites; (D) iHS with 80% segregating sites.

### **2.3.3 Relationship between False Discovery Rate and power to detect sweeps**

We investigated the relationship between iHS power and false discovery rate (FDR) under different combinations of varying selective strengths and extent of missing data. iHS power was calculated as the average power for sweep frequencies ranging from 0.1 to 1.0. With complete data, iHS achieved greater than 80% power while keeping false discovery rate at 1% or lower under strong selection. It still retained 60% power under medium or weak selection. Under all selective strengths tested, the iHS power for any given sweep scenario (partial vs. complete sweeps) for complete data was reduced by about 20% when analyzing datasets that are missing 40% of the actual segregating sites (Fig. 2.3). Overall, iHS power starts to plateau with false discovery rate as low as about 1%, and further increasing FDR does not result in remarkable improvement over iHS sensitivity.

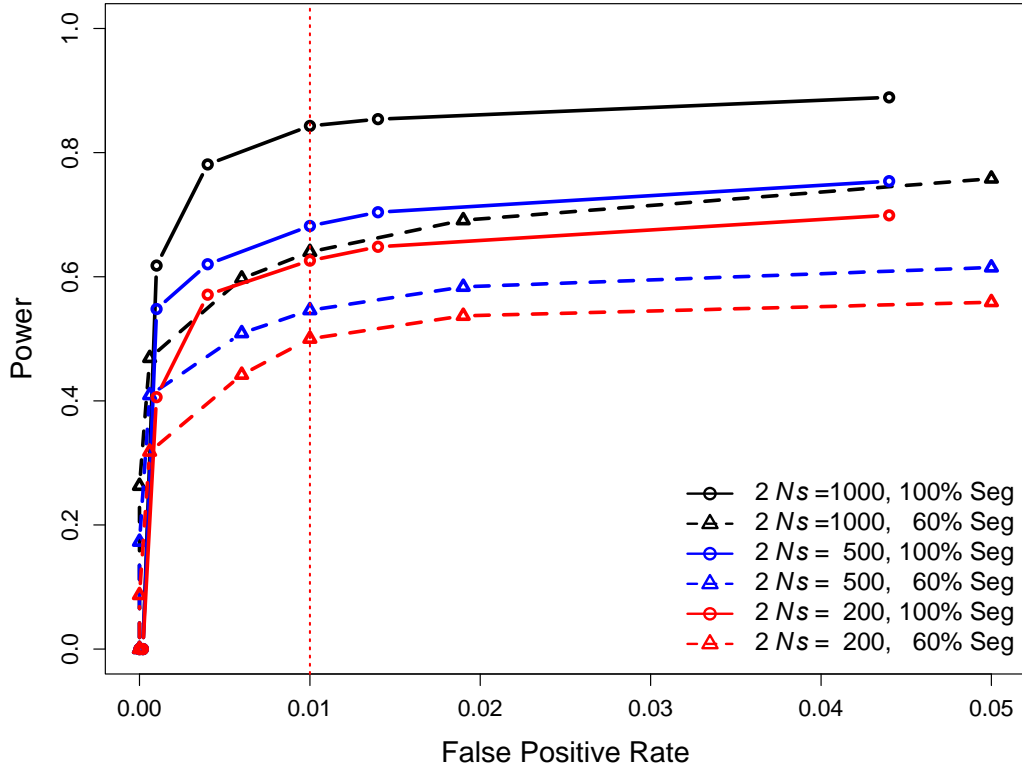


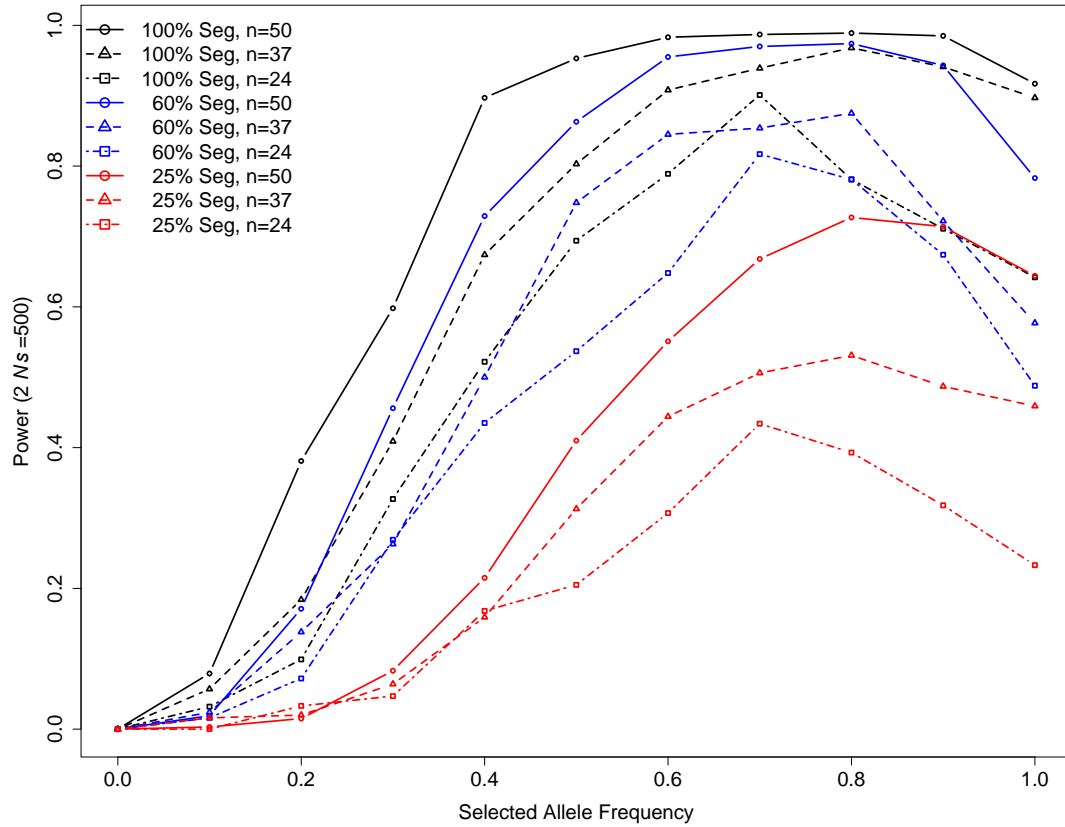
Figure 2.3: Relationship between iHS power and false discovery rate.

Power of iHS was estimated under each false discovery rate given using population samples of 50 individuals. Complete data (100% Segregating sites, solid lines) were compared with low-quality datasets (60% Segregating sites, dashed lines) for varying selection scenarios (strong selection,  $2N_e s=1000$ , black; medium selection,  $2N_e s=500$ , blue; weak selection,  $2N_e s=200$ , red). Power was calculated as the arithmetic average of powers for all selected allele frequencies in the interval of 0.1 to 1.0. Vertical dashed line indicates the FDR value used in most analysis in this paper (FDR=1%).

#### 2.3.4 Effect of sample size on low-quality datasets

Low-pass sequencing of multiple samples is one of the typical experimental designs in many population surveys nowadays. Therefore, we assessed the benefits of larger sample size to identify sweeps using iHS on datasets with different qualities. Our results showed (Fig. 2.4) that under medium selection, increasing sample size had more pronounced improvement for mid- to high-frequency sweeps, but not for low-frequency ones. This finding was also confirmed with strong and weak selection (data not shown).

Within each level of missing data (100%, 60% or 25%), larger sample size improves sensitivity. However, we note that the level of missing data has an even more detrimental effect on iHS sensitivity. Specifically, for sweeps with frequencies greater than 0.5, increasing sample size from 24 to 37 chromosomes enhanced iHS power by 9.3%, 14.8%, 16.6% and from 37 to 50 by 4.2%, 15.0% and 17.5% when 100%, 60% and 25% of total segregating sites available in the data, respectively (Fig. 2.4). Thus, data completeness enhances iHS sensitivity greater than increasing sample size. Keeping data completeness constant, sampling more individuals is more important to datasets with lower quality than to those with relatively higher quality, especially in identifying mid- to high-frequency sweeps.



**Figure 2.4: Effect of sample size and missing data on the power of iHS to detect recent selective sweeps.**

Power was calculated for simulations with  $2Nes = 500$  and sample sizes of 50 chromosomes (solid lines), 37 chromosomes (dashed lines) or 24 chromosomes (dotted lines). Three scenarios of missing data were simulated for each of these sample sizes: complete data (black), 60% segregating sites (blue), 25% segregating sites (red).

## 2.4 DISCUSSION

We evaluated the performance of two commonly used DNA polymorphism-based methods (SweepFinder and iHS) to detect recent selective sweeps when the datasets contain missing base calls and sequence errors. We first assessed two different simulation schemes (short-read emulation and random missing variants) for generating sequence datasets typical of low quality Illumina sequences in the case of both neutrality and with complete and partial selective sweeps and we saw quite comparable results (Fig. 2.1).

The SweepFinder method was developed to use full sequence site frequency spectrum data to evaluate the fit to a neutral model compared to a model with a recent complete selective sweep (KIM and STEPHAN 2002). As expected, we found that it had little to no power to detect partial (incomplete) sweeps. We also found that SweepFinder was very sensitive to missing data and sequence errors and its power was greatly compromised with low-quality datasets. In contrast, the haplotype-based method iHS exhibited high sensitivity to both recent partial and complete sweeps, and strong robustness to missing data and sequence errors when retaining very low false discovery rate. In addition to data quality issues, SweepFinder would further lose power due to its sensitivity to demography (JENSEN *et al.* 2007b) while iHS remains robust to demographic assumptions (NIELSEN *et al.* 2007; SABETI *et al.* 2007).

We also contrasted the robustness of the two methods to the strength of selection ( $2N_e s$ ). Both methods are more powerful and robust to missing data and sequence errors when datasets are generated under strong selection ( $2N_e s=1000$ ), while footprints from weak selection are poorly detected with datasets containing missing data and sequence errors. Consistent with previous findings (PAVLIDIS *et al.* 2010), SweepFinder power is more sensitive to stronger selection compared to iHS. We also noticed that iHS was able to detect the majority of putative sweeps predicted by SweepFinder. This is because SweepFinder has high power to detect recent strong selective sweeps, which also tend to create extended haplotype structure in the data. Such haplotype signals are readily detected by iHS, and thus most discoveries made by SweepFinder are recovered by iHS as well. With weak selective sweeps and missing data and sequence errors, SweepFinder has much lower power while iHS still retains much of its power. Impressively, even with weak selection and with 40% of the actual segregating sites missing, iHS was able to identify about 80% of the mid- to high-frequency end-point putative partial sweeps with very low false discoveries (FDR=1%). In contrast, for these latter conditions, SweepFinder had very low power, only identifying about 2% of medium frequency and 20.7% of fixed sweeps.

Under weak selection, many sweeps detected by iHS were not detected by SweepFinder. This contrast is likely due to biases in the allele frequency spectrum that result from an inflated variance in allele frequency in low-coverage sequencing datasets



(LYNCH 2008). In particular, rare variants are under-sampled in low-pass datasets (JIANG *et al.* 2009; LYNCH 2009; CRAWFORD and LAZZARO 2012), which biases the allele frequency spectrum and consequently reduces SweepFinder power. In contrast, the under-representation of rare variants only slightly impacts iHS power, presumably because they contribute little to haplotype structure.

As a technical aside, inferences of selection from SweepFinder and iHS analyses of complete sweeps might be less powerful than expected by the number of chromosomes sampled alone when the data are obtained from heterozygous diploid individuals. Unequal sampling of the two alleles in heterozygous individuals by next-generation sequencing could lead to an excess of rare variants in the polymorphism data (HELLMANN *et al.* 2008; JOHNSON and SLATKIN 2008; LYNCH 2008; JIANG *et al.* 2009; KIM *et al.* 2011), which is a signature of recent complete sweeps and thus can lead to false predictions of recent positive selection. Therefore, the homozygosity we assumed in generating short-reads may result in smaller but more unbiased estimate of power on complete sweeps. Furthermore, phasing is not involved in this study because homozygous/inbred individuals were simulated. However, next gen sequencing of heterozygous diploid individuals and associated errors in phasing inference may lower the sensitivity of haplotype-based methods.

While iHS is more robust to missing data and high-throughput errors, it typically generates a large confidence interval in locating sweeps (GROSSMAN *et al.* 2010). We

observed that SweepFinder was able to locate sweep targets more accurately than iHS when complete data was used. However, when applied on datasets with missing data, the variance of SweepFinder prediction accuracy increased dramatically; in contrast, iHS prediction accuracy was barely affected. Stronger sweeps enhances SweepFinder accuracy dramatically as expected. There is, however, an inverse relationship between the strength of selection and iHS prediction accuracy likely due to the fact that stronger selection drives faster fixation of the beneficial mutations and leaves longer haplotype footprints. Under the most extreme case, very strong recent selection wipes out all variants nearby and iHS would only be able to pick up signals distant away from the selected locus where linkage disequilibrium is deteriorated. Thus, to locate targets of strong sweeps with high-quality data, SweepFinder has better resolution, while iHS is more appropriate for examining targets of weak sweeps or with low-quality data. One possible way to narrow down the targets of putative sweeps might be to take advantage of composite statistics that utilize information from several weakly correlated statistics (GROSSMAN *et al.* 2010).

We also assessed how sample size, missing data and sequence errors in combination would affect the performance of iHS in finding recent sweeps. With complete, high quality data simulated with population parameters appropriate for *Drosophila melanogaster*, signals of selection could be efficiently detected by iHS with relatively small sample size of only 24 chromosomes. However, detection of partial

sweeps of low frequency in the population did not benefit much in the power gain by increasing sample size. Our analysis also suggests that larger sample sizes but reduced data quality has only marginal and diminishing improvement on the power to identify recent sweeps. Missing data in the sequences appears to be more detrimental to power and outweighs the power gain from larger sample size. Calling SNPs individually or from reads pool of all samples may lead to different number of SNPs being called and may improve the power of both methods. Thus, larger sample size may improve the probability of telling true variant sites from invariant ones, but it does not improve the quality of each single base call.

In summary, the power to detect recent selective sweeps using next-gen population sequence datasets can be severely limited by reduced data quality and incompleteness. A major problem is that many statistical methods rely on the full site frequency spectrum for every site, which is not available for most next-gen sequence datasets. Thus while whole-genome next-gen sequencing scans can achieve large sample sizes, missing data as well as errors in SNP calling are likely to undermine such power gain, especially under low-pass experimental design or in non-model organisms without good reference genomes. Fortunately, haplotype-based approaches remain relatively robust to reduced site coverage and data quality in both their power to detect and to localize recent selective sweeps. New statistical approaches incorporating both site-frequency spectrum and haplotype information have been recently developed

(GROSSMAN *et al.* 2010; PAVLIDIS *et al.* 2010) may show better performance over traditional composite likelihood models such as SweepFinder (Jeffrey D. Jensen, personal communications) and it will be worth exploring how such methods would perform on low-quality next-gen datasets with sequencing errors and missing data.

## **2.5 ACKNOWLEDGMENTS**

This work was supported in part by a National Institutes of Health (R01-GM36431 to C.F.A). The authors thank Andrew G. Clark, Jeffrey Jensen, Haley Hunter-Zinck and Alon Keinan for helpful discussion and feedback on this work and manuscript.

## **Chapter 3. Computational Inference of selective sweeps in a 2 Mbp region in *Drosophila melanogaster* X chromosome**

### **3.1 INTRODUCTION**

Population geneticists have long been exploring how natural selection shapes genetic variation. When a new beneficial mutation is selected, its frequency will elevate within the population and bring other variations nearby to higher frequency or fixation as well (SMITH and HAIGH 1974; BEGUN and AQUADRO 1992). This pattern is typically referred as selective sweeps, which sweeps out variations in nearby region if the process finishes relatively fast. Recent sweeps typically skew SFS towards an excess of both low- and high-frequency derived mutations and leave footprints of reduced level of variation. They will also change the haplotype structure of the region and lead to unusually long segments of increased linkage disequilibrium (LD) (STEPHAN *et al.* 2006; PAVLIDIS *et al.* 2010). As a result, there have been many new computational approaches developed in the last decade or so to identify either recent or recurrent selective sweeps.

As mentioned in Chapter 1, these methods can be typically classified into two large categories: divergence-based and polymorphism based. Polymorphism-based methods can be further divided by site-frequency spectrum (SFS)-based and linkage disequilibrium (LD)-based (NIELSEN *et al.* 2007). Briefly, between-species comparisons are employed to identify older and recurrent events, while polymorphism-based

methods are used for more recent events of selection. Some methods compare statistics against neutrality derived from certain demographic models (WILLIAMSON *et al.* 2005; LI and STEPHAN 2006; KEIGHTLEY and EYRE-WALKER 2007), while some others utilize genome-wide SNP frequency distributions (NIELSEN *et al.* 2005b). Outlier-based approaches are also widely used in many population genetic investigations: briefly, one or more commonly used statistics are computed for all marks in the entire dataset and a top fraction are deemed as an indicator of departure from population genetic equilibrium expectations, thus corresponding regions are considered candidates targets of selection (OLEKSYK *et al.* 2008).

In population genetic surveys, especially at genomic scale, discovery of the same selected gene regions by multiple computational approaches is typically deemed as strong evidence for selective sweeps in the region. However, we should also be aware that, contradicting results from two methods, i.e. the success of one test and the failure of the other, does not exclude selection in the region because of many factors, including the fact that different methods may be designed for different types of sweeps in different population periods, and there are always limitations on sensitivity (SABETI *et al.* 2006; KELLEY and SWANSON 2008).

We here focus on identifying regions of recurrent and recent positive selection within a 2 Mbp region of *Drosophila melanogaster* X chromosome using five different computational approaches: MK-test (CHARLESWORTH and EYRE-WALKER 2008),

SweepFinder (NIELSEN *et al.* 2005b), SweeD (PAVLIDIS *et al.* 2013), Omega (PAVLIDIS *et al.* 2010), and iHS (VOIGHT *et al.* 2006). By comparing and contrasting results from each method, we further want to discuss both the advantages and limitations of each method and hope this will help people when evaluating and interpreting each statistic, so that people are aware of the importance of choosing proper statistic for a particular purpose of population genetic investigations.

## **3.2 MATERIALS AND METHODS**

### **3.2.1 *Drosophila melanogaster* chromosome X 2 Mb resequencing data**

We used the same 20 lines used in the Singh *et al.* (2013) population survey, which included 20 chromosome X genomic re-sequencing data of *Drosophila melanogaster* sampled from a single population in Namulonge, Uganda (POOL and AQUADRO 2006; SINGH *et al.* 2013). The re-sequencing data targeted a 2 Mb region of X chromosome between gene *garnet* and *scalloped*, with coordinates from 13,621,236 to 15,719,755 in FlyBase release 5. Each X chromosome was assured to be isogenic by the chromosome extraction procedure. The samples were sequenced using Illumina single-end technique with reads of 86 bp and had coverage in our target 2 Mb region ranging from 16.9× to 67.4×. The 20 Uganda samples were aligned with *Drosophila melanogaster* genome FlyBase release 5.2 using BWA and SNPs were called by the joint genotype for inbred lines (JGIL) (STONE 2012).

### 3.2.2 Computational approaches

The McDonald–Kreitman (MK) test was used to identify recurrent positive selection. It compares regional polymorphism level with divergence level at both neutral and functional sites and uses a simple but elegant 2x2 test of independence to evaluate possible deviation from neutrality expectations. The MK test is generally robust to variation of mutation rate across the genome but may be confounded by slightly deleterious mutations or nonequilibrium demography. It is one of the most popular approaches and has been widely applied across both model and non-model organisms to infer recurrent adaptive fixations.

To detect recent selective sweeps, we employed both SFS-based methods (SweepFinder, SweeD) as well as LD-based methods (iHS, Omega) based on their popularities in current population genetic studies.

SweepFinder utilizes the genomic site-frequency spectrum (SFS) information to compute the likelihood of observing the allele frequency at each SNP position and compare the total likelihood within a sliding window to the neutral assumption to infer selections. This approach originated from the composite-likelihood ratio test (CLRT) developed by Kim and Stephan (2002), but the SFS of the null hypothesis is derived from the background SFS rather than the standard neutral model. From this substitution, people argue that SweepFinder is more robust to demography and mutation rate variation in the genome. SweepFinder was initially designed to identify completed



sweeps at a genomic scale, and has very poor power in detecting partial sweeps as shown in Chapter 2 even with complete data.

SweeD is a likelihood-based improvement over SweepFinder. It is capable of calculating neutral SFS for given demographic parameterization without the need to empirically compute the genomic SFS as the background (PAVLIDIS *et al.* 2013) (<http://sco.h-its.org/exelixis/software.html>). It is claimed to be able to analyze very large datasets of thousands of sample sequences. The program also offers an option to include monomorphic sites that help sweep detection in low SNP density regions. In contrast, SweepFinder typically skips calculations for such regions of low SNP density (PAVLIDIS *et al.* 2010).

Polymorphisms on either side of a target of a selective sweep evolve independently with recombination. Thus there predicted to be significant LD within each side but not across the sweep target (STEPHAN *et al.* 2006). This pattern in the genome can be detected by the  $\omega$ -statistic, which measures the extent of LD within each side of the target of the sweep but not across the sweep target (KIM and NIELSEN 2004). Omega is a variable-window size modification of the  $\omega$ -statistic that was designed to deal with variable recombination rates and recurrent sweeps in nearby region (PAVLIDIS *et al.* 2010).

The LD-based method, iHS, uses regional haplotype structure to make predictions about selective sweeps. An EHH score is calculated for each individual SNP in the dataset, which has higher value with stronger LD and longer haplotypes (SABETI *et al.*

2002). Then it integrates the EHH scores over genetic map and compares the LD decay of the derived with the ancestral alleles at the same SNP site, which cancels out the influence of local variable recombination rate. iHS is more appropriate for analyzing large datasets since large collection of SNP's are needed to normalize the raw iHS scores for comparisons. Though it was initially designed to detect partial sweeps, it also has high power in predicting complete sweeps as pointed out in Chapter 2.

### **3.2.3 Determining significance of the statistics**

The popular coalescent simulator ms was used to simulate neutral models with the -s option to match the number of segregating sites and SNP density (HUDSON 2002). Over 10,000 neutral replicates were simulated independently under the Wright-Fisher equilibrium model. DnaSP v5.0 was used to perform the McDonald-Kreitman test for each specific gene (<http://www.ub.edu/dnasp/>).

Significance of SweepFinder, SweeD and Omega inference results were determined by essentially following the same SweepFinder procedure as described in Chapter 2. Briefly, the critical value was calculated from the empirical distribution consisting of likelihood ratios from the ten thousand neutral simulations generated above given FDR=1%, and a sweep is predicted if the likelihood ratio of the sample is greater than the critical value determined.

The iHS significance evaluation followed what was described in *Materials and Methods* of Chapter 2. In summary, SNPs under a sweep model were deemed to be significant if its iHS value fell outside of the 99% confidence interval, which was estimated from the empirical distribution of the ten thousand iHS values calculated from the neutral simulations generated above given FDR=1%. A putative target of selective sweep was considered statistically significant if more than 30% of the SNPs in a 10-SNP sliding window (with step size of 5 SNPs) had significant iHS scores.

### **3.3 RESULTS**

For the 2.1 Mb region analyzed, 21.35% of total nucleotide sites contain at least one missing base call across the 20 individuals, and 11.65% sites have more than three base calls missing. To retain only high-quality nucleotide sites and compare how the four computational approaches would perform on sequenced data with missing base calls, all following analysis only include sites that have complete data across all 20 individuals.

The nine targets of selection were predicted by all four methods in the nine-gene region located between coordinates 15,600-15,642 kbp of the X chromosome. This region has a very low level of missing data (only 2.76% of total base pairs are called N's in the region) and very strong signals of selection. In addition, both region 13,922-13,924 kbp and 15,227-15,280 kbp were only identified as targets of recent sweeps by Omega and iHS, but not by SweepFinder or SweeD (Table 3.1). This difference may be partly

because these two regions have much higher extent of missing data (21.64% sites missing for 13,922-13,924 kbp with maximum of 9 individuals missing at the same sites; 16.53% sites missing for 15,227-15,280 kbp with maximum of all 20 individuals missing at the same sites). As indicated in Chapter 2, SFS-based approaches are more sensitive to data completeness and higher level of missing data will reduce their power pronouncedly, while haplotype-based approaches are overall more robust to missing data.

Interestingly, 15,227-15,280 kbp accommodates 6 genes while region 13,922-13,924 kbp has no currently known or predicted genes, thus might be selection acting on regulatory elements of genes outside of this latter region. By performing a sliding 10-kb window independent of annotation and number of segregating sites included, all three regions, namely 13,922-13,924 kbp, 15,227-15,280 kbp and 15,600-15,642 kbp, exhibited dramatic reduction in polymorphism level measured by Watterson's Theta and Pi (coordinates on the physical map: 15,226,236–15,301,236 and 15,596,236, 15,636,236; Fig. 3.1), which suggests possible footprints of recent selective sweeps.

For this 2.1 Mbp region examined, SweepFinder and SweeD both were able to identify 13 putative genes (9 genes in the 15,600-15,642 kbp region) showing significant signs of being positively selected in recently population evolution (Table 3.1). Though sharing the same set of top putative targets, SweeD showed slightly more significance in p-value of rejecting neutrality, mostly likely because it has reduced sensitivity to low

SNP density due to the fact that the method also incorporates monomorphic sites in analysis to improve performance especially in regions with fewer number of SNPs (CRISCI *et al.* 2013).

In contrast, LD-based methods demonstrated higher sensitivity. Omega showed slightly higher sensitivity to complete sweeps by detecting 16 genes in total and a region with no known genes (13,922-13,924 kbp), all of which showed the most extent of reduction in polymorphisms (Fig. 3.1). However, Omega missed four top predictions made by SweepFinder and SweeD, out of which three are also predicted by iHS (Table 3.1). iHS was found to have the highest power to identify recent selection, including both complete and partial sweeps. It recovers all findings of Omega's, three out of four findings specifically made by SweepFinder and SweeD, as well as its own several novel findings.

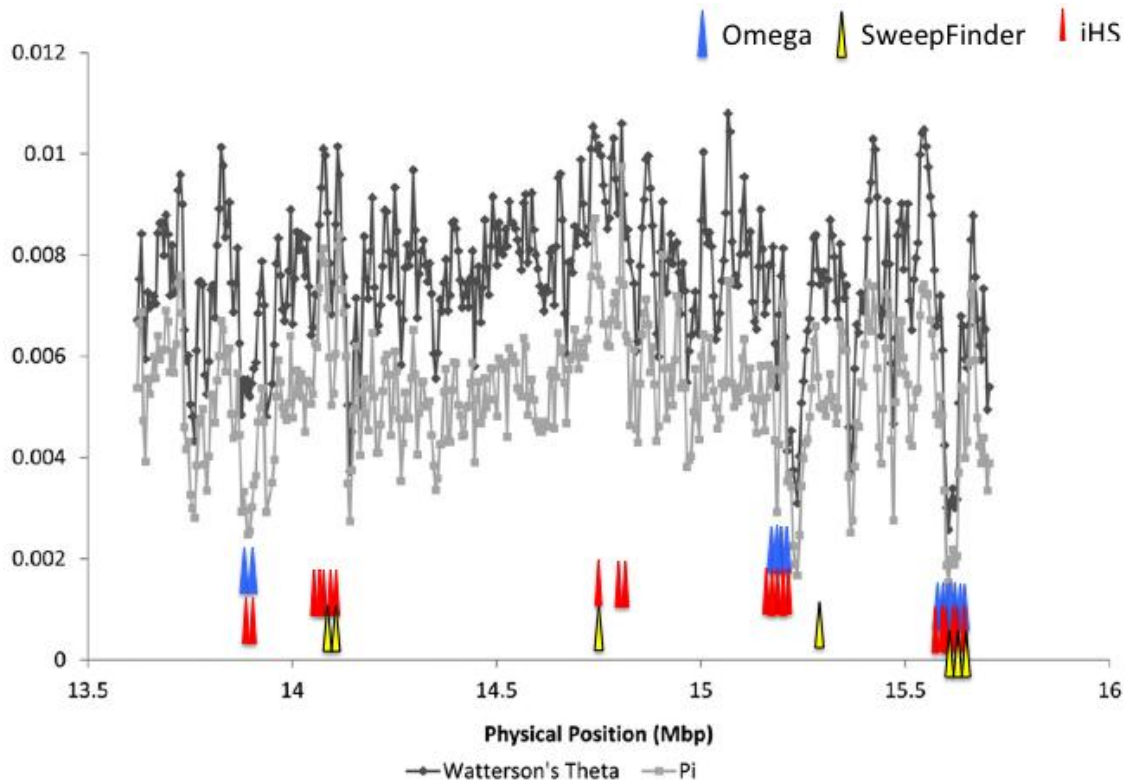


Figure 3.1: Top putative targets of recent positive selection inferred by various computational approaches.

(modified from Singh, *et al*, 2013). Top putative targets predicted by each method is labeled in corresponding colored triangles: Omega (blue), SweepFinder (yellow), iHS (red). SweepD had the same set of top predictions as SweepFinder and is not listed here separately.

**Table 3.1 Top putative regions or genes identified to be targets of recent selective sweeps by each method.**

There are multiple genes within coordinates 15,227-15,280 kbp (6 genes) and 15,600-15,642 kbp (10 genes), respectively, and all the genes are inferred to be under recent adaptive selection if the specified method made the predictions.

Gene (kbp)	Omega	SweepFinder	iHS	SweeD
13,922–13,924	YES		YES	
15,227–15,270*	YES		YES	
15,600–15,642*	YES	YES	YES	YES
CG9203		YES		YES
<i>Fbxl4</i>		YES	YES	YES
<i>mud</i>		YES	YES	YES
Flo-2 (intron)		YES	YES	YES
yl			YES	
Pdgy			YES	
CG32604			YES	
CG13403			YES	
CG11674			YES	

In addition to the 15 genes residing in 15,227-15,280 kbp and 15,600-15,642 kbp, there are three genes showing very significant iHS scores, and we performed McDonald-Kreitman tests to test for recurrent sweeps at these top putative targets. Different from the pairwise alignments between *D. melanogaster* and *D. sechellia* used in Singh *et al* (2013), we employed a more rigorous three-species alignment of *D. melanogaster*, *D. sechellia* and *D. yakuba* and inferred the ancestral state of each nucleotide site. Sites whose ancestral states could not be inferred were excluded from the MK-test and the *D. melanogaster* and *D. sechellia* pair alone are examined. Out of the 15 genes found predicted to be targets of selection by at least two methods, only two are significant after Bonferroni correct for multiple testing. Only two of the three genes solely predicted by iHS to have had selective sweeps showed a significance departure from neutrality with the MK-test (*Flo-2* and *fbxl4*) (Table 3.2). This result could be conservative compared to that without ancestral state inference in detecting recurrent sweeps since some informative sites are not included in the analysis.



**Table 3.2 Results of McDonald-Kreitman test of top putative genes.**

The first 15 genes residing in 15,227-15,280 kbp and 15,600-15,642 kbp and were predicted by at least three out of the four methods. The last three genes were the top putative targets identified solely by iHS. MK-test column shows the *p*-value before multiple-testing correction and Bonferroni *p*-value column is after. Significant *p*-values (<0.05) after multiple-testing correction are labeled in red, and non-significant but *p*-values < 0.5 are in green.

	Gene (kbp)	MK-test	Bonferroni p-value
Omega, SweepFinder, SweeD, iHS	FBgn0030630c	4E-4	0.008
	FBgn0030631c	0.31	1
	acj6	0.47	1
	FBgn0030672de	0.33	1
	FBgn0030673e	0.15	1
	FBgn0030674e	1E-5	2E-4
	FBgn0030675e	0.42	1
	FBgn0030676e	0.35	1
	actr13E	0.92	1
	FBgn0033391e	0.02	0.35
	l(1)G0136	1	1
	FBgn0030678d	0.087	1
	FBgn0030680c	0.01	0.18
	FBgn0030628c	0.83	1
	HDAC6	0.68	1
iHS	Flo-2	0.004	0.08
	Pdgy	0.07	1
	Fbxl4	0.001	0.022

### 3.4 DISCUSSION

We applied four different computational approaches, two SFS-based and two LD-based, to detect both complete and partial selective sweeps, on a 2.1 Mb segment of *D. melanogaster* X chromosome. The three regions on the physical map that have been identified in previous studies, namely 13.922–13.924 Mb ( $p < 0.0001$ ), 15.248–15.250 Mb ( $p < 0.0001$ ), and 15.600–15.602 Mb ( $p < 0.0001$ ), are also recovered in our studies, and more novel putative targets were identified by iHS. Although evolutionary forces other than recent selection (e.g. non-equilibrium demography) must be considered as alternative explanations to the departures from neutrality observed, non-equilibrium demography alone is not sufficient to explain the observed selection inference (SINGH *et al.* 2013).

Of the four statistical tests we applied to this 2.1 Mb region of *D. melanogaster*, iHS detected the largest number of putative sweeps (though it did not infer selection at CG9203). Since iHS is reasonably robust to demographic assumptions and generally has lower false discovery rate compare with other three methods for single hitchhiking events (SABETI *et al.* 2007; CRISCI *et al.* 2013), the higher number of predicted targets of selection is likely due to the fact that iHS is less sensitive to the scattered missing data and sequence errors in this next-gen dataset. It remains possible, however, that iHS predictions have a higher proportion of false discoveries if most of the segment had experienced both recurrent sweeps and non-equilibrium demography, which will lead to higher false positive rate than SweepFinder and SweeD (CRISCI *et al.* 2013).

There are arguments about how sensitive LD-based approaches are to non-equilibrium demography. It was previously shown that iHS is generally robust to various population bottlenecks by using human parameterization (SABETI *et al.* 2007; GROSSMAN *et al.* 2010), while there was also evidence showing that both Omega and iHS performance might be affected by non-equilibrium demographic assumptions and would lead to higher false discoveries (CRISCI *et al.* 2013). The discrepancy between these two sets of studies can be reconciled by the fact that the former two studies only simulated single selective sweeps under low to intermediate level of population bottleneck scenarios for individual selective events, while the latter study simulated both single sweeps as well as recurrent sweeps with full-range of population bottleneck severities for both single hitch hiking and recurrent hitch hiking occurrence. Generally, when a population bottleneck is of low to intermediate severity, LD-based approaches are still robust and only show slight increase in false discoveries. But if the bottleneck is severe, LD-based methods still demonstrated slightly lower rate of false discovery compared to SFS-based methods in detecting individual selective sweeps. For recurrent sweeps, Omega showed much higher false positive rate under non-equilibrium demography. However, Omega was designed to detect the LD footprints left by recent complete sweeps, while iHS is more sensitive to partial sweeps. The false positive rate of iHS in the presence of recurrent sweeps remains to be evaluated.

The iHS method predicted multiple genes that might have experienced recent selective sweep among the 2.1 Mb segment. Out of the three top putative target genes, *Flo-2*, *pdgy* and *fbxl4*, two also showed signatures of recurrent sweeps by the MK test. Interestingly, *Drosophila Flo-2* gene has been reported to be a required component in restricting the spread of epidermal wound response (JUAREZ *et al.* 2011) and is an important player of the morphogens *Wnt* and *Hedgehog* (KATANAEV *et al.* 2008), and was also inferred independently to be the target of a recent selective sweep (WERZNER *et al.* 2013). The *fbxl4* gene was shown to be significantly associated with immunization in human (LI *et al.* 2010b). Also, the 3'-end of *pdgy* gene is also very close (~1 kb) to the non-coding DNA that was reported to be one of the outliers among the ~250 loci surveyed in African *D. melanogaster* (OMETTO *et al.* 2005).

Both SFS-based methods SweepFinder and SweeD demonstrated very similar performance and also slightly underperformed the other LD-based approach Omega (13 genes by SweepFinder and SweeD vs. 15 genes by Omega). This difference is likely due to the fact that SFS-based methods are more sensitive to incomplete data and performance will be compromised (JENSEN *et al.* 2007a; SINGH *et al.* 2013) (Chapter 2). In comparison, Omega tends to perform more conservatively since it failed to detect selection for the four genes that are top putative targets of all other three methods. In addition, only two out of the 15 genes predicted by these three methods had significant MK test results, compared to two of three for iHS, which suggests methods designed to

identify complete sweeps may overall be more sensitive to missing data issue, even though Omega also utilizes haplotype information to make inferences. Note that significance of a MK test for a particular gene might reasonably be viewed as an indicator that the gene truly did experience recent selective sweeps. In contrast, a lack of significant departure with the MK test only means that there is no evidence of repeated selective sweeps in the past for this gene, and is not evidence against a single recent selective sweep inferred by the other methods.

In summary, our study emphasized the relative advantages of LD-based methods in detecting recent sweeps relative to SFS-based approaches when applied to incomplete data typical of next-gen sequence polymorphism surveys. Combine the higher sensitivity of LD-based approaches and the higher resolution in locating sweeps provided by SFS-base approaches, approaches that use a composite of both SFS- and LD-based methods might be valuable (ZENG *et al.* 2007; GROSSMAN *et al.* 2010). However, how such composite methods will perform under large datasets with missing data and sequence errors is an area that needs further investigation. Therefore, researchers need to be careful to choose the methodology that best fits their specific goals. In addition, there remains a clear need for future method development designed specifically for incomplete datasets.

## **Chapter 4. An HMM-based program for peak detection and estimation of transcript boundaries from high-throughput transcriptome sequencing data**

### **4.1 INTRODUCTION**

RNA-seq is rapidly becoming a popular technique for studies of gene expression (HILLIER *et al.* 2009; WANG *et al.* 2009; NAGALAKSHMI *et al.* 2010), with transcriptionally active regions (TARs) inferred from localized enrichment of sequencing reads (a.k.a. “peaks”) in sequence alignments. Current analysis pipelines have many limitations. For instance they cannot predict and define transcription starting sites (TSS) and transcription ending sites (TES) without existing annotation models, or simply use arbitrary read-count cutoffs with maxGap/minRun segmentation (KAMPA *et al.* 2004; ROYCE *et al.* 2005; TRAPNELL *et al.* 2010; HABEGGER *et al.* 2011), and confidence is usually estimated by FDR in permutation-based ways (MORTAZAVI *et al.* 2008; PARK 2009; PEPKE *et al.* 2009).

Here, we present a flexible statistical program, HPIBD (HMM-based Peak Identification and Boundary Definition) for *de novo* analysis of RNA-seq datasets. It is based on a three-layer HMM model taking into account of effects of local GC content. It avoids the use of arbitrary read-depth cutoffs and has built-in tolerance to read gaps. It is able to process RNA-seq in a strand-specific way, make strand-specific peak

predictions and statistically evaluate transcript boundaries (TSS and TES). We implemented the model and showed HPIBD has robust performance under various validations and with benchmark to Cufflinks.

## **4.2 MATERIALS AND METHODS**

### **4.2.1 HPIBD framework**

HPIBD partitions the genome into transcribed and un-transcribed segments based on the number of sequencing reads mapped to each nucleotide site. And the program takes five major steps to statistically infer whether a site is transcribed or not.

In the preprocessing step, local GC content of each site is calculated from the reference genome. Then the local GC content is converted to one of the binary states: high GC region or low GC region. Both the window size that HPIBD needs to look at to calculate local GC content and the threshold separating high GC regions and low GC regions can be tuned by users for different research goals with various organisms.

Secondly, HPIBD reads in the read depth at each nucleotide site, in Pileup format (LI 2011). The model is set up differently from a typical multi-state Hidden Markov Model (HMM) in that it adds dependency of read depth and transcript state on local GC content, as well as constraints on minimum peak length and minimum gap length, the latter of which are tunable to users for various organisms. Then the model is initialized

with random seeds that are generated internally, and makes an initial guess of dataset-specific model parameterizations.

In the next step, HPIBD applies an unsupervised learning technique called Baum-Welch algorithm on the modified multi-state HMM to optimize the dataset-specific model parameterizations from the initial guess iteratively. The estimated parameters for the model are optimized specifically to the input dataset when maximum likelihood is gained after several iterations.

In the fourth step, the optimized dataset-specific model parameters are applied to statistically infer the posterior probability of the transcription state of each site under the maximum likelihood principle using Viterbi algorithm.

Lastly, with default or user-defined confidence threshold, a list of transcript states of all sites is generated based on the posterior probabilities calculated from the previous step.

The time complexity of the program is  $O(LM^2I)$ , where  $L$  is the length of the input sequence in bare pair,  $M$  is the number of possible peak states ( $M=P+B$ ; where  $P$  is min peak length, and  $B$  is min gap length),  $I$  is the number of iterations for the program to converge. Empirically, the value of  $I$  depends heavily on data quality, and poor quality data typically needs more iterations to find the optimal solution.



#### 4.2.2 Notations

Let  $L$  be the length of input reference genome sequence in bp for analysis. Large reference genomes can be split into multiple smaller segments to be analyzed individually in parallel. In the latter case, let  $L$  be the length of a specific small segment.

Let  $U$  be the local GC content, and  $u_i \in U$  is the local GC content at a specific nucleotide position  $i$ , where  $i$  ranges from 1 to  $L$ . Each  $u_i$  can be individually calculated from the reference genome given position  $i$  and local GC window size  $w$  (default = 1,000 bp) and is defined to be the ratio of G and C nucleotide counts over the GC window size  $w$ . Thus,  $u_i$  is modeled to be a discrete distribution over  $[0, 1]$ . For each nucleotide position  $i$ , read depth is also known and noted as  $x_i$ . (Fig. 4.1A)

Let  $Z$  represent the transcription state of the nucleotides. Assume  $z_i \in Z$  indicates the transcription status of a specific nucleotide site, where  $z_i=1$  means the site is transcribed in the dataset, and  $z_i=0$  means the site is not transcribed.

Let  $P$  be the minimum peak length in bp and  $B$  be the minimum gap length in bp between adjacent peaks. Then  $M$  is the number of possible peak states in the model where  $M=P+B$ .  $I$  is the number of iterations before the program converges to an optimal parameterization of the entire model and start to make transcription status inference for each nucleotide.

### 4.2.3 Statistically Modeling

We applied a three-layer multi-state HMM to model the transcription status of each nucleotide (Fig. 4.1A). Each nucleotide in the reference sequence is classified into either transcribed (peak status) or not transcribed (background status). The classification is not binary, but instead, a posterior probability ( $p$ ) for each nucleotide site belonging to the peak status is calculated so that users can adjust the threshold to be either more conservative or more aggressive in a particular project. The default setting for the posterior probability is that, a site is evaluated to be transcribed/expressed if it is more likely to be in peak status than in background status ( $p>0.5$ ).

We employ a three-layer multi-state HMM because of the following reasons: 1. There is strong correlation of transcription status between adjacent nucleotides. That is, a nucleotide in peak status is more likely next to another nucleotide also in peak status, and vice versa; 2. There is strong correlation of read depth between adjacent nucleotides; 3. The multi-state setup controls for false peak and background inference and allows more flexibility for different species; 4. Whether the nucleotide at position  $i$  is within a peak (transcribed,  $z_i=1$ ) or not (non-transcribed,  $z_i=0$ ) is dependent on local GC content. This dependency comes from the fact that high GC-content regions tend to be coding, thus more likely to be transcribed than low GC-content ones.

The local GC-content dependence is modeled as a Bernoulli distribution with  $p$  value being the threshold user specified or by program default.

$$sgn(u_i) = \begin{cases} 0, & \text{if } u_i < GC \text{ threshold} \\ 1, & \text{if } u_i \geq GC \text{ threshold} \end{cases}$$

The transition probabilities are designed in the way that it is 1 within minimum peak length and within minimum background length, but beyond that, it follows a Bernoulli distribution between peak status and background status, depending on local GC content, as the following (Fig. 4.1B).

$$p(z_i = 0, z_{i+1} = 0 | u_i) = \lambda_{sgn(u_i)}$$

$$p(z_i = 0, z_{i+1} = 1 | u_i) = 1 - \lambda_{sgn(u_i)}$$

$$p(z_i = 1, z_{i+1} = 1 | u_i) = \eta_{sgn(u_i)}$$

$$p(z_i = 1, z_{i+1} = 0 | u_i) = 1 - \eta_{sgn(u_i)}$$

To infer the emission probabilities in the HMM, we used Gaussian Tail distribution (GT) and Gamma distribution to model read depth at non-transcribed and transcribed sites, respectively:

$$p(x_i | z_i = 0, u_i) \sim GT(x \geq 0, \sigma^2)$$

$$p(x_i | z_i = 1, u_i) \sim Gamma(k_{sgn(u_i)}, \theta_{sgn(u_i)})$$

There is significant read depth over-dispersion (the variance of read depth is much larger than the mean read depth among sites) in the RNA-seq datasets. Poisson distribution requires equal mean and variance, which is not appropriate for this type of data. In contrast, Gamma distribution is adopted to accommodate the over-dispersion by optimizing the combination of the shape and scale parameters.

An HMM is designed to perform inference on each nucleotide and a posterior marginal probability that the nucleotide is transcribed (a.k.a. in peak status) is calculated for final inference of transcribed regions.

#### **4.2.4 Model parameterization optimization and estimation**

We employed the well-established Baum-Welch algorithm to perform dataset-specific optimization and estimation of HMM model parameters. Observed transition and emission probabilities were summarized as prior inputs of the model to speed up optimization. An iterative approach was adopted to optimize parameters: within each round of optimization, the posterior probabilities of all hidden states (transcribed, or non-transcribed) for each nucleotide are derived, conditioning on the current estimates of model parameterization. Then conditional on the inferred hidden states, model parameters are updated sequentially using the method of moments. The iteration proceeds until the likelihood of observing the data given the model parameterization converges and reaches maximum.

The dataset-specifically optimized model parameterization is then used to infer the transcription status of each nucleotide using Viterbi algorithm. The posterior probability of each single nucleotide being transcribed is calculated and boundaries of the transcribed regions were then identified using either user-specified or default posterior probability threshold and outputted.

#### 4.2.5 Flexibility

Users can set minimum peak length that is expected in the species under analysis or some other closely-related species. Minimum gap length could also be set to compensate for abnormally large read coverage gaps in the data. By default, unit size  $z_i$  is set to be 1bp in size to give the maximum resolution in defining transcription boundaries. But use can specify a larger unit size (e.g. 100 bp) for much faster speed if identification of peaks is the primary purpose of the analysis. There are other options user can fine tune to achieve better performance for specific species, such as the local GC-content window size.

#### 4.2.6 Implementation

We implemented HPIBD in a C program, which can run on most platforms that support C language compilation and execution, e.g. Unix/Linux, Windows, Mac, et al. On a workstation computer powered by Intel Core i7-860 (8M Cache, 2.80 GHz) and 16GB RAM, HPIBD usually analyzes a 1Mb region in 3~5 min and uses reasonable amount of memory (~630 MB). Different chromosomes and regions can be analyzed in parallel, making it easily applicable to genome-wide data. The *Drosophila melanogaster* chromosome X is about 22 Mbp long and can be analyzed in approximately 20 min by breaking up the X chromosome into four equally long segments of 5.5 Mbp and

analyzing on four CPU cores in parallel, or about 80 min if running on a single CPU core, with approximately 14 GB memory usage in both cases.

#### **4.2.7 Benchmarking with Cufflinks**

Cufflinks was downloaded from <http://cufflinks.cbc.umd.edu>. The specific version used for benchmarking is 2.0.2 release. No prior annotations were used in our inferences from either Cufflinks or our HPIBD method. The same min intron length of 50 bp was used for both Cufflinks and HPIBD, which is a bit conservative for *Drosophila melanogaster* (PRESGRAVES 2006). While Cufflinks does not offer such an option, HPIBD also specified a minimum peak length of 50 bp.

#### **4.2.8 *Drosophila melanogaster* Tiling Array data and RNA-seq data**

Tiling array data were downloaded from the modENCODE database for *Drosophila melanogaster*. RNA expression profiling data from embryo 0-2 hour stage and 10-12 hour stage were collected (<http://data.modencode.org>; Author: Celniker S; Total-RNA; 38 bp resolution; DCCid: modENCODE\_101, modENCODE\_102, modENCODE\_105, modENCODE\_106). Thirty-two regions of varying sizes were analyzed and used out of all potential regions that were significantly higher expressed in the embryo 0-2 hour stage by comparing the two stages. RNA-seq for embryo 0-2 hour stage of *Drosophila melanogaster* were also downloaded from modENCODE

(<http://data.modencode.org>; Author: Graveley B; Poly-RNA; 76 bp; DCCid: modENCODE\_4439). RNA-seq data of *Drosophila melanogaster* S2-DRSC cell line was used for benchmarking HPIBD with Cufflinks and was downloaded from modENCODE (<http://data.modencode.org>; Author: Graveley B; Total-RNA; 38 bp; DCCid: modENCODE\_983). Pileup files for benchmarking were generated from SAM files downloads using the mpileup command in SAMtools (LI 2011).

#### **4.2.9 Differential expression analysis with expression profiling tiling array data**

We used edgeR in Bioconductor package to evaluate differential expression from tiling array data (ROBINSON *et al.* 2010). The R package was downloaded from <http://www.bioconductor.org/packages/2.12/bioc/html/edgeR.html>

#### **4.2.10 *Drosophila melanogaster* reference genome, FlyBase annotation, and EST data**

Both the reference genome of *Drosophila melanogaster* and the annotation were downloaded from FlyBase release v5.3 (MARYGOLD *et al.* 2013). And the EST data was downloaded from Drosophila Gold Collection of Berkley Drosophila Genome Project ([http://www.fruitfly.org/EST/gold\\_collection.shtml](http://www.fruitfly.org/EST/gold_collection.shtml))

#### 4.2.11 Simulation of RNA-seq data

Transcription active regions (TARs) and non-transcribed regions (NTRs) are simulated in alternating order, the lengths of which follow exponential distributions with means of 1,000 bp and 5,000 bp, respectively. Total of two scenarios were simulated: “medium expression, medium noise”, and “low expression, high noise”.

In medium expression, medium noise scenario, each nucleotide in the reference genome was simulated independently, and simulating each nucleotide to be a G or C follows Bernoulli distribution with  $p=0.8$  if within TARs and  $p=0.4$  if within NTRs, otherwise the nucleotide is randomly chosen from A or T with equal probability. We simulated read depth every 10 bp and all nucleotides within the same 10 bp window have the same read depth coverage. This is to reflect both the correlation between adjacent base pairs as well as varying read depth. The read depth for each 10 bp bin was independently drawn from Gamma distributions and rounded to nearest integers for both TARs and NTRs. Specifically, read depth follows  $\text{Gamma}(k=5, \theta=15)$  at TARs, and  $\text{Gamma}(k=0.5, \theta=5)$  at NTRs, in which mean read depth of TARs is 75 reads and that of noise is 2.5 reads.

In low expression, high noise scenario, we followed very similar procedures in medium expression, medium noise scenario. But we modeled each nucleotide to be a G or C follows Bernoulli distribution with  $p=0.6$  if within TARs and  $p=0.5$  if within NTRs. Another modification is that read depth follows  $\text{Gamma}(k=4, \theta=5)$  at TARs, and

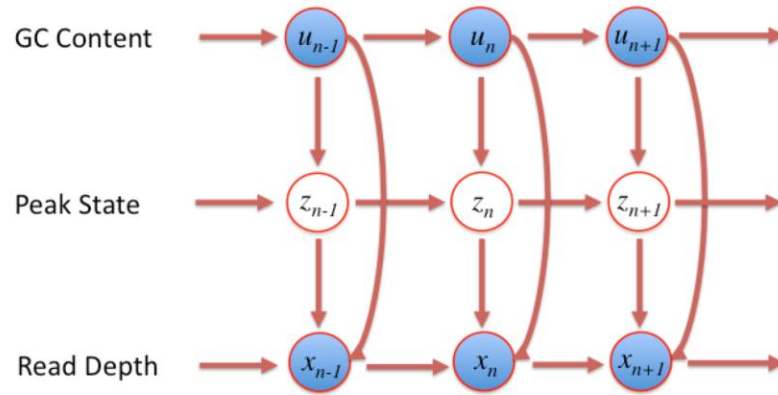


Gamma( $k=0.5$ ,  $\theta=15$ ) at NTRs, in which mean read depth of TARs is 20 reads and that of noise is 7.5 reads with much larger variance.

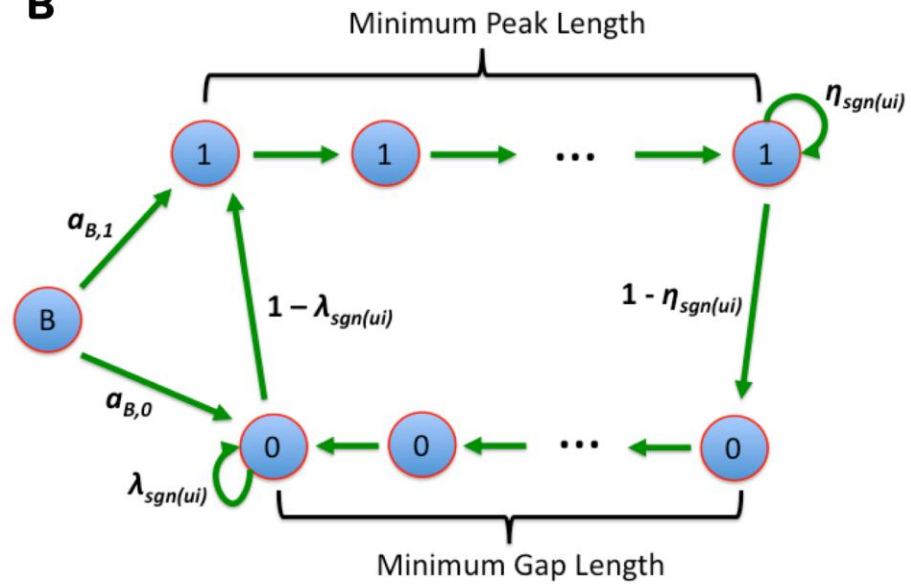
**Figure 4.1: Graphic representation of HPIBD.**

(A) Three layer of HMM was designed and dependencies are indicated by red arrows. Filled circles are observed variables and empty ones are latent variables. (B) Transition map between peak states. Algorithm enters peak states at first position of the sequence at probability set by the users (default,  $a_{B,1}=a_{B,0}=0.5$ ). Only peaks longer than the minimum peak and gaps longer than minimum gap length would be predicted. Transition within the minimum peak length is set to be one.

**A**



**B**



## 4.3 RESULTS

We applied HPIBD to *Drosophila melanogaster* RNA-seq data and compared its results against tiling array, EST data, and FlyBase annotation. We also benchmarked HPIBD with the most popular RNA-seq analysis program Cufflinks on both simulated and experimental datasets. The performances of the programs were evaluated mainly from three aspects: (1) sensitivity to identify TARs; (2) false positive rate; and (3) accuracy in defining transcript boundaries.

### 4.3.1 HPIBD performance against tiling array results

As an initial step, we evaluated both the sensitivity of HPIBD to identify TARs and its accuracy in defining boundaries of TARs. We first evaluated the expression level at each embryo stage and identified 32 regions of varying sizes that are expressed significantly higher in the embryo 0-2 hour stage compare to that in embryo 10-12 stage between chromosome 2L coordinate 0 bp and 1 Mb. These 32 regions are therefore used as “truly” transcribed regions. HPIBD was then applied on the same chromosome 2L region of *Drosophila melanogaster* RNA-seq data at embryo 0-2 hour stage, and results were compared against those from tiling array analysis. All 32 “truly” transcribed regions identified by tiling array analysis were successfully recovered by HPIBD RNA-seq analysis. Further all HPIBD defined transcript ends agreed with that from tiling array analysis, and fell within the tiling array resolution of 38 bp.

#### 4.3.2 HPIBD performance against EST data and Fly annotations

For the 32 TARs identified above, HPIBD results were highly consistent with both EST data and FlyBase annotations. Specifically for transcript boundary definition, 31 out of the 32 regions had no discordance among HPIBD, EST and FlyBase models. Only 1 region showed 1 bp discordance at 3'-UTR between HPIBD and the other two models.

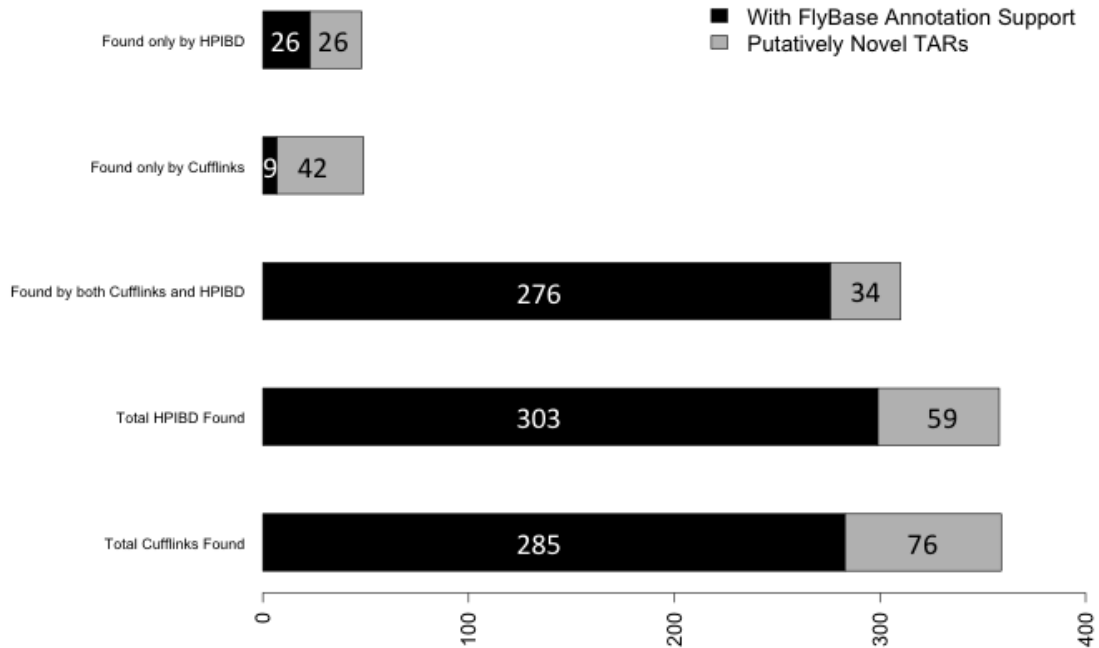
#### 4.3.3 Benchmark of HPIBD with Cufflinks on experimental RNA-seq data

We benchmarked HPIBD performance with Cufflinks using *Drosophila melanogaster* S2-DRSC cell line RNA-seq data. We compared the results from a randomly picked 1 Mb region of *Drosophila melanogaster* chromosome 2L (coordinates: 0 bp–1 Mb) with each other program as well as FlyBase annotations. All inferences of TARs were identified *de novo* and no annotation files were provided for either program.

We applied both Cufflinks and HPIBD on the 1Mb region and first evaluated their capability of identifying TARs *de novo*. The results showed that Cufflinks and HPIBD identified about the same number of TARs (Cufflinks: 361 TARs, HPIBD: 362 TARs, Fig. 4.2). Comparing their findings with FlyBase annotations, HPIBD might have slightly higher sensitivity by having 303 TARs confirmed by annotations, while Cufflinks had 285 TARs consistent with annotations.

As for potential false discoveries, 21.1% of total TARs found by Cufflinks were not confirmed by annotations, while only 16.3% by HPIBD were not. In contrast, out of the 310 TARs that were found by both programs, only 11.0% could not be confirmed by FlyBase annotations. Thus, the lower proportion of annotation-not-confirmed inferences suggests that HPIBD has a lower false discovery rate than Cufflinks. This is also confirmed by the fact that only 9 out of 51 TARs (17.6%) were confirmed by annotations for TARs that were inferred only by Cufflinks; however, 26 out of 52 TARs (50.0%) identified only by HPIBD were supported with annotation evidence (Fig. 4.2).

There are 42 insertions that showed signature of transcription within the 1 Mb alignment, and Cufflinks failed to identify 7 of them. At the same time, HPIBD detected them all, demonstrating strong robustness to potential transcribed insertions in alignments. In addition, there are two instances where FlyBase annotates as two adjacent exons, while Cufflinks made inference of one linked exon, and HPIBD was able to match the FlyBase gene model. On the other hand, there are two large exons having greater than 60 bp read coverage gap that HPIBD erroneously inferred to be two exons, respectively. Cufflinks was able to make annotation-consistent inference on one exon, while failed to identify the second.



**Figure 4.2: Comparison of Cufflinks and HPIBD on sensitivity of finding TARs.**

A 1Mb region on chr2L from *Drosophila melanogaster* was randomly chosen and results from both programs were analyzed against FlyBase annotations.

We also evaluated the accuracy of each program in defining TAR 5' and 3' boundaries. The 276 TARs that were found by both Cufflinks and HPIBD were used in this analysis, and their transcription starting sites (TSS) and transcription ending sites (TES) were obtained using FlyBase annotations as the gold standards for evaluation. The TSS was further manually corrected with TSS refined mapping data (NECHAEV *et al.* 2010) for better reliability. We define the accuracy being the distance between the true TSS/TES and the program predictions.

For 5'-UTR definition, Cufflinks tends to have a systematic bias of defining TSS towards more downstream to the true TSS. With the TARs surveyed, Cufflinks defines TSS with the median accuracy of about 18 bp downstream. Interestingly, HPIBD has a positively skewed distance distribution of TSS accuracy. Most TSS predicted by HPIBD fell very close to the true TSS, but there is also non-trivial proportion that fell downstream, while the median accuracy of HPIBD is 1 bp upstream (Fig. 4.3). Further, we used coefficient of variation (CV) to measure which programs predictions are closer to annotated TSS, and the CV of Cufflinks to HPIBD is 1.17 for TSS. These results suggest that Cufflinks slightly underperforms HPIBD in defining 5'-UTR TSS and tend to define TSS towards downstream.

For 3'-UTR definition, Cufflinks had slight tendency of defining TES downstream and HPIBD slightly tend to be more upstream. But both program exhibited no obvious systematic biases in accuracy and showed reasonably symmetric distance



distribution of accuracy with near 0 bp median (Fig. 4.4). The variance of distance increase dramatically for Cufflinks in contrast to slight increase for HPIBD, which suggests both programs perform poorer in defining TES but Cufflinks was much more pronouncedly affected. The coefficient of variation (CV) of Cufflinks to HPIBD is 1.76 for TES. These together strongly suggest that HPIBD outperforms Cufflinks substantially in accurately defining 3'-UTR TES.

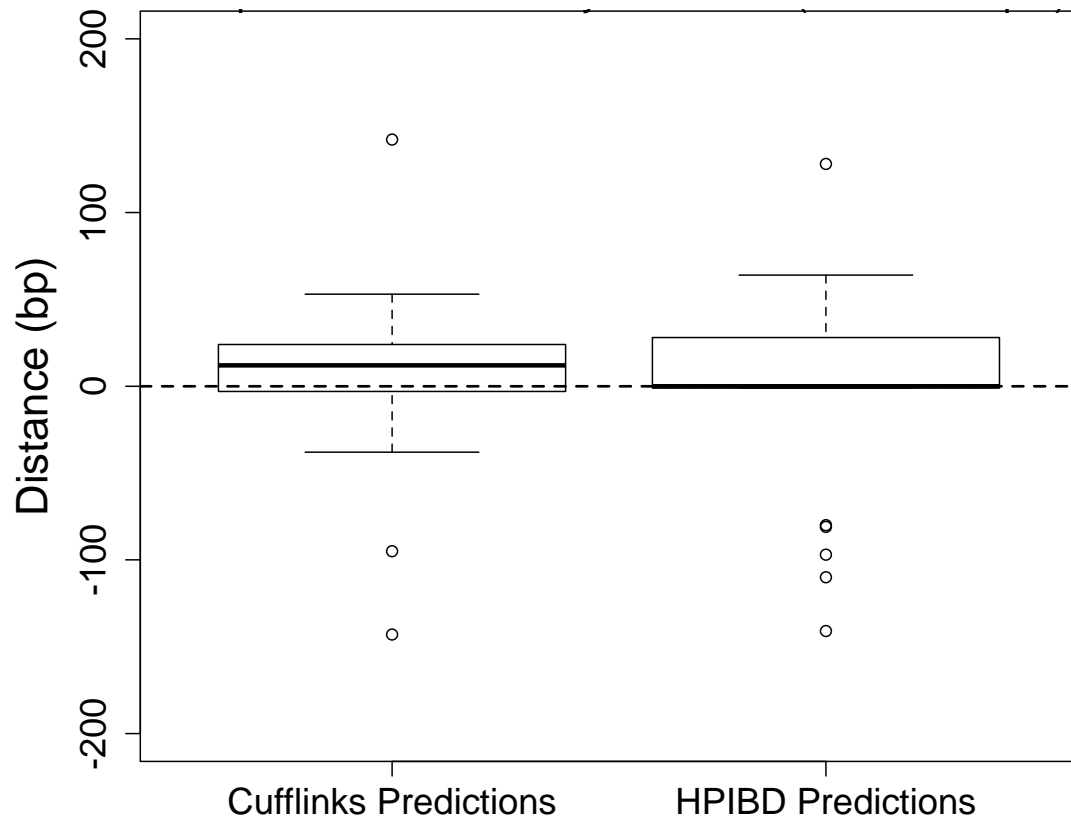
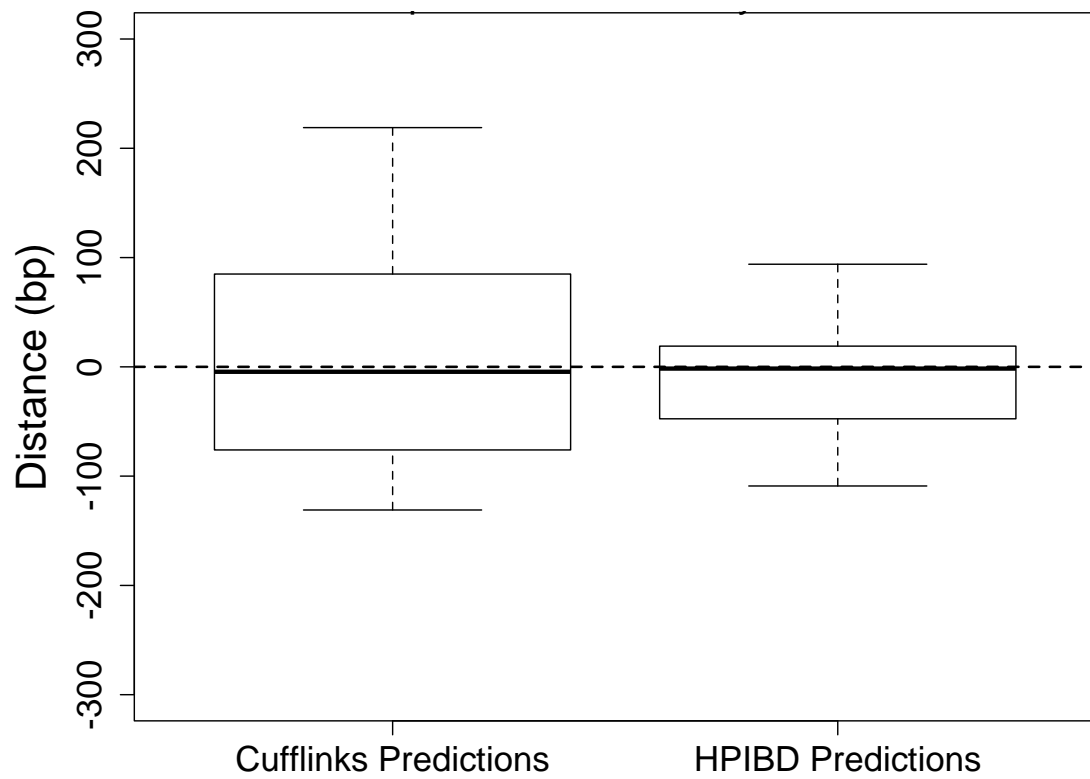


Figure 4.3: Distance between predicted and annotated TSS at 5'-UTR.

Annotation transcription starting sites (TSS) were downloaded from FlyBase and manually correct with (NECHAEV *et al.* 2010).



**Figure 4.4: Distance between predicted and annotated TES at 3'-UTR.**

Annotation transcription ending sites (TES) were downloaded from FlyBase.

#### 4.3.4 Benchmark of HPIBD with Cufflinks on simulated RNA-seq data

Simulation studies were also carried out to produce two different types of RNA-seq data: “medium expression, medium noise” and “low expression, high noise”. The former was done with the purpose of imitating general medium to low quality RNA-seq datasets, which have deep read coverage but also noticeable uniformity in sequencing read-depth. The latter scenario examines how Cufflinks and HPIBD perform under stress conditions where there are dramatic sequencing background noises around lowly expressed genes.

Under medium expression, medium noise scenario, a total of 492 TARs was simulated with the same number of NTRs (Table 1). Cufflinks inferred 516 TARs, which was more than the number of true TARs, indicating there was a 6% false discovery rate among its inferences. In contrast, HPIBD did not make any false inference in this simulation. Both programs had very high sensitivity and were able to recover about 99% of true TARs with the simulated high expression level and medium noise.

HPIBD slightly outperformed in defining both TSS and TES correctly (HPIBD 97% vs. Cufflinks 89%). Also, Cufflinks could locate TSS and TES as far as 17 bp or 14 bp off the true sites, but almost all HPIBD predictions of TSS and TES (99%) fell within  $\pm 1$  bp of the true sites, compared to that of 91% inferred by Cufflinks.

As expected, when 467 TARs were simulated under the much more stressed “low expression, high noise” scenario, both programs showed compromised performance. Cufflinks had only 67% sensitivity but mistakenly included pronounced proportion (47%) of false discoveries in its inferences. In contrast, HPIBD retained reasonably high sensitivity of 88% while keeping the false discoveries very low at only 2%.

Cufflinks was able to locate only 12% of TSS and TES without errors for TARs that it correctly identified, while HPIBD had a much higher proportion (40%) of its predictions correctly. Further, only 21% of Cufflinks defined TSS and TES were both within  $\pm 1$  bp from the true sites, whereas 77% of those defined by HPIBD fell in the same accuracy. What is to note is, both programs failed to accurately define the TES of the same TAR due to the high noise and high non-uniformity of the read depth across sites, thus leading to an erroneous inference of 460 bp from the true site. Excluding this outlier TAR, HPIBD consistently showed better accuracy in inferring TES than Cufflinks.

**Table 4.1 Simulation evaluations of HPIBD performance.**

TARs and non-transcribed regions (NTRs) are simulated in alternating order, and details in *Materials and Methods*. Completely Correct TARs refer to TARs that were successfully identified, and the TSS and TES of which were correctly defined by the programs.

Expression Noise		Medium Medium	Low High
Total TARs No.		492	467
Identified TARs	Cufflinks	516	596
	HPIBD	490	416
Number of False Positives (False discovery rate)	Cufflinks	31 (6%)	283 (47%)
	HPIBD	0 (0%)	7 (2%)
Number of True Positives (Sensitivity)	Cufflinks	485 (98.6%)	313 (67%)
	HPIBD	490 (99.6%)	409 (88%)
Completely Correct TAR (% correctly predicted TARs)	Cufflinks	433 (89%)	37 (12%)
	HPIBD	475 (97%)	164 (40%)
Max Frame Error <sup>§</sup>	Cufflinks	31 bp	460 bp (57 bp) <sup>¶</sup>
	HPIBD	2 bp	460 bp (11 bp) <sup>¶</sup>
Max TSS Error	Cufflinks	17 bp	38 bp
	HPIBD	1 bp	10 bp
Max TES Error	Cufflinks	14 bp	460 bp (23 bp) <sup>¶</sup>
	HPIBD	1 bp	460 bp (7 bp) <sup>¶</sup>
TSS within $\pm 1$ bp (% correctly predicted TARs)	Cufflinks	434 (89%)	62 (20%)
	HPIBD	481 (98%)	224 (55%)
TES within $\pm 1$ bp (% correctly predicted TARs)	Cufflinks	436 (90%)	65 (21%)
	HPIBD	482 (98%)	232 (57%)
Both Ends within $\pm 1$ bp (% total true TARs)	Cufflinks	440 (91%)	66 (21%)
	HPIBD	488 (99%)	313 (77%)

**§ Maximum TAR boundary shift difference between inferred TARs and corresponding true TARs.**

**¶ There was only one predicted TAR had ending position 460 bp off. Number in parenthesis showed results excluding that TAR.**

## **4.4 DISCUSSION**

We have implemented an HMM-based statistical method for peak finding and TAR boundary definition without the limitations of using arbitrary read depth cutoffs. This flexible program employs the Expectation-Maximization algorithm and is tolerant to variation in sequencing depth of RNA-seq data. Each TAR is statistically evaluated and strand-specific data are readily fit into the framework.

One notable feature that the program (HPIBD) provides is the scalability to large datasets. And there are two techniques HPIBD supports for faster analysis. The first technique is parallel analysis. Users can analyze different chromosomes in parallel or break down a large chromosome into smaller segments at sites with no reads around, e.g. large intergenic regions, et al, and analyze separately or in parallel.

The other technique is to employ a user-specified unit option. Given the same minimum peak length and minimum gap length, memory usage increases linearly with the length of the sequence to be analyzed, while time cost increases slightly faster than linear due to the fact that longer sequences may lead to more iterations to optimize the model parameterization. By default, HPIBD runs at a resolution of 1 bp, which approximately takes an hour and a half to finish analyzing a 20 Mbp region. Running at resolution of 10 bp will speed up the analysis to about 10 times faster (~7 minutes), and



user can later focus and re-analyze putative transcript regions in higher resolution of 1 bp.

We evaluated HPIBD performance from three aspects: sensitivity, false positive rate, and accuracy in boundary definition. By comparing HPIBD results with results from tiling array, EST models and FlyBase annotations, we found that *de novo* analysis of RNA-seq data by HPIBD had very high sensitivity while accurately defined transcript boundaries with minimum errors.

We further benchmarked HPIBD performance with Cufflinks on a 1 Mb region using *Drosophila melanogaster* S2-DRSC RNA-seq data and compared inferences with FlyBase annotations. Results suggested HPIBD had slightly higher sensitivity than Cufflinks due to better concordance with annotations. Further, fewer HPIBD inferences was novel to the comprehensive FlyBase annotations, which strongly suggests HPIBD has higher specificity, thus lower false discovery rate compared to Cufflinks.

It is possible that some putative findings that are not confirmed by FlyBase annotations are truly novel TARs, e.g. small RNAs, since the dataset was from Total-RNA preparation, and about 11.0% of common findings by both programs were not documented in FlyBase annotations. But given the comprehensiveness of FlyBase annotations and the fact that Cufflinks had much higher proportion of putative findings that were not existed in annotations (82.4% in those found only by Cufflinks, 50.0% in

those found only by HPIBD, compared to 11.0% in those found by both programs), it is very much likely the difference in results was due to Cufflinks underperforming HPIBD in identifying TARs.

Both programs showed similar accuracy in defining TSS, but it seemed that Cufflinks has a systematic bias of defining TSS downstream of true sites; while HPIBD, on the other hand, exhibits much lesser biases. However, HPIBD demonstrated dramatically higher accuracy in defining TES comparing with Cufflinks, and both programs showed little or no systematic biases.

The simulation studies showed that under normal conditions of medium expression and medium noise, Cufflinks and HPIBD had comparable sensitivity in identifying TARs, and HPIBD outperformed slightly in both specificity (lower false discovery rate) and accuracy in defining TSS and TES sites. However, with low-quality RNA-seq datasets of high noise level and the purpose of identifying lowly expressed regions, the performance of Cufflinks was dramatically compromised in both identifying TARs and locating their TSS and TES. In contrast, HPIBD exhibited strong robustness to such low-quality scenario and retained much of its power and ability to accurately infer TSS and TES.

The program would be particularly useful when there is no prior knowledge about gene models in susceptible regions/genomes, e.g. non-model organisms where

comprehensive and accurate annotations are not available yet. HPIBD is designed for true transcription detection in relatively noisy datasets. Further, the very high resolution of the program makes it applicable to a wide variety of potential applications, such as studying transcription frame shifts under different conditions with RNA-seq data and to study transcription rate with GRO-seq data (CORE *et al.* 2008). Furthermore, due to the flexibility in the model setup, HPIBD might also be applicable to CHIP-seq datasets to detect enriched regions of transcription factor binding sites.

## **Appendix A: Analysis of DPGP V1 and V2 datasets reveals numerous sequencing biases and errors**

### **A.1 INTRODUCTION**

The short-read large-scale sequencing is the most popular approach to collect massive amount of data at declining cost in nearly all fields of current biological studies (see (ZHANG *et al.* 2011) for methodology review). While new technologies are been developed and tested (e.g. single-molecule sequencing), Illumina sequencing (BENTLEY *et al.* 2008), 454 Life Sciences (Roche) pyrosequencing (MARGULIES *et al.* 2005), Applied Biosystems SOLiD sequencing (MARDIS 2008) have been the recent the dominant sequencing platforms available on the current commercial market. The potential applications of next-generation (NGS, also called second-generation) sequencing techniques have been drastically strengthened by the more accurate whole genome assemblies in model organisms and tremendous improvement for non-model organisms (e.g. (DONIGER *et al.* 2008; OSSOWSKI *et al.* 2008; WANG *et al.* 2008; WHEELER *et al.* 2008; AHN *et al.* 2009; HILLIER *et al.* 2009; GENOMES PROJECT *et al.* 2010; SCHWARTZ *et al.* 2010; GENOMES PROJECT *et al.* 2012; BRADNAM *et al.* 2013). Such assemblies, which are still under fast improvements, serve as references for short reads mapping, and various downstream bioinformatic analysis, including genetic variation discovery, RNA

expression analysis, DNA-protein interaction and epigenetic surveys (see (KOBOLDT *et al.* 2013; RIVERA and REN 2013) for review).

One major application of NGS is variant detection. The advantage of variant identification by sequencing is that most variants, common or rare, known or novel, nucleotide or structural, can be discovered with corresponding sequencing methodology, sequencing depth and coverage and appropriate bioinformatic software. As previously mentioned, a reliable reference genome serves as a starting point, so that various variant calling algorithms can be employed for downstream applications such as population genetic surveys (e.g. SNP calling (LI *et al.* 2008a; LI *et al.* 2008b; KOBOLDT *et al.* 2009; LANGMEAD *et al.* 2009; LI *et al.* 2009b; MCKENNA *et al.* 2010; SHEN *et al.* 2010; DEPRISTO *et al.* 2011; LIU *et al.* 2012); indel detection (YE *et al.* 2009; EMDE *et al.* 2012; ONMUS-LEONE *et al.* 2013), etc.).

Due to the complex nature in the chemistry and following analysis, errors and missing data could stem from any of the steps. Though typically sequencing accuracy can be improved by incorporating more individuals (larger sample size) (GENOMES PROJECT *et al.* 2010; GENOMES PROJECT *et al.* 2012; LANGLEY *et al.* 2012; MACKAY *et al.* 2012; POOL *et al.* 2012), extending the coverage of the genome (deep sequencing) (e.g. (BENTLEY *et al.* 2008), and applying advanced bioinformatic algorithms (e.g. analyze all reads from all samples together instead of individual-specific base calling) (STONE 2012), there is always trade-off between quantity and quality, especially for organisms with

large genome size. Under the given budget and specific research goals, sequencing coverage and sample size are always evaluated against the statistical power and errors that will be used for lower-coverage datasets, and such trade-off should be taken into account in the experimental design. Given the large size and potentially low quality/coverage of the data with regard to individual calls for any individual base across all lines, it is necessary to explore which methods can still be used to identify genome-scale selected regions.

## **A.2 MATERIALS AND METHODS**

### **A.2.1 Sanger Sequences of *Drosophila melanogaster***

Fifteen individuals of *Drosophila melanogaster* were sequenced by Sanger sequencing technique for a 22 kb region around the *Notch* gene on X chromosome. All 15 individuals come from California, USA and sequences were generated in our lab(DUMONT *et al.* 2004).

### **A.2.2 DPGP samples**

Whole genomes re-sequencing assemblies of 37 inbred lines of *Drosophila melanogaster* from North Carolina, US and 117 *D. melanogaster* individuals from 20 populations in Africa, were downloaded from the Drosophila Population Genomics Project (DPGP) website (<http://www.dpgp.org>).

The sequenced North American *D. melanogaster* genomes of DPGP V1 employed first generation (single-end 36 bp) Solexa/Illumina technology(BENTLEY *et al.* 2008) and was assembled using MAQ 0.6.8(LI *et al.* 2008a). The sample consists of 37 inbred genomes from Trudy Mackay's set of inbred lines sampled in Raleigh, NC(JORDAN *et al.* 2007) and a set of sequenced chromosomes (7 chrX's, 6 chr2's and 5 chr3's) from a sample of Malawi isofemale lines(BEGUN and LINDFORS 2005) that were inbred using balancers. It is claimed that the unique portions of the *D. melanogaster* genome had coverage of greater than 10X. Regions of repeated sequence are filtered in the release (set to "N"). In comparison with the Sanger/ABI data, the same 22 kb region were extracted for analysis with in-house Perl scripts.

The 117 individuals of DPGP V2 were largely from sub-Saharan Africa with one individual from Lyon, France. All individuals were sequenced using the haploid embryos technique(LANGLEY *et al.* 2012) with Illumina GAIIx (75bp, paired end) technology. Sequencing errors and biases had been extensively modeled and corrected by the DPGP panel using the reference genome and FASTQ releases were downloaded for this analysis.

### **A.2.3 SweepFinder and iHS Calculation and Sweep Evaluation**

We used SweepFinder (NIELSEN *et al.* 2005b) to perform the Composite Likelihood Ratio Test (CLRT). SweepFinder takes the polymorphism data and compares

the likelihood of the model with a sweep to the likelihood of the neutral model without selection but based on genome background allele frequencies. A grid size of 10, equivalent to 1 kb, was used in the analysis. The critical value was calculated from the empirical distribution consisting of likelihood ratios from the ten thousand neutral simulations generated above given FDR=1%. Statistical significance was then evaluated and a sweep was “detected” if the likelihood ratio of the sample was greater than the critical value determined as described above.

To confirm the simulated samples under selection contained sufficient SNPs for SweepFinder to construct background allele frequencies, we performed two different sweep detection and significance evaluation procedures after matching SNP density of simulated samples to Drosophila Population Genomics Project Illumina dataset (release 1.0 for *D. melanogaster*; <http://www.dpgp.org/>). In the first approach, we applied SweepFinder on sweep simulated and edited samples, and constructed background site frequency spectrum (SFS) directly from neutral SNPs in the sample sequences. Alternatively, we used neutrally simulated samples with all of the same parameters (except that  $s=0$ ) to construct the background allele frequencies for SweepFinder. We obtained identical inferences of selection using both sets of neutral reference SNPs.

The iHS program was downloaded from Pritchard lab site (<http://hgdp.uchicago.edu/>). iHS utilizes haplotype structure to examine the existence of extended haplotypes on the genetic map (VOIGHT *et al.* 2006). The 99% confidence



interval was estimated from the empirical distribution of the ten thousand iHS values calculated from the neutral simulations generated above given FDR=1%. SNPs under a sweep model were deemed to be significant if its iHS value fell outside of the 99% confidence interval. A putative target of selective sweep was considered statistically significant if more than 30% of the SNPs in a 10-SNP sliding window (with step size of 5 SNPs) had significant iHS scores. Power was estimated as the proportion of identified sweeps to the total number of simulated sweeps (calculated from a minimum of 100 simulations). Predicted target sites of selection were defined to be the midpoint of the significant 10-SNP windows (VOIGHT *et al.* 2006; GROSSMAN *et al.* 2010).

### **A.3 RESULTS**

#### **A.3.1 Sequencing errors and biases in DPGP V1 compared with Sanger sequences**

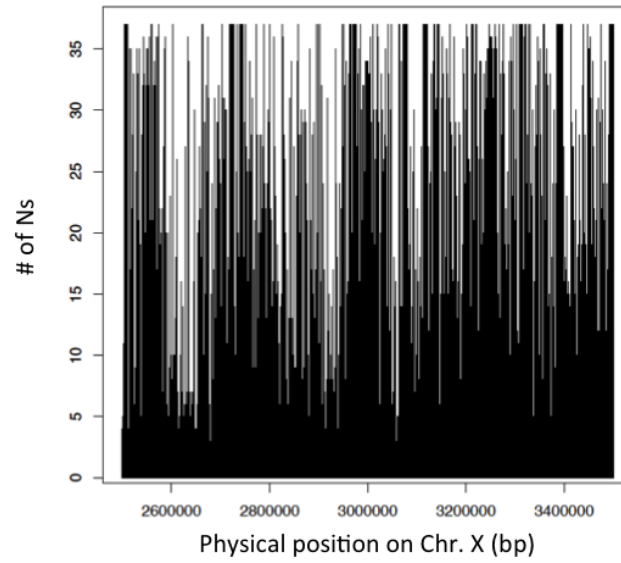
We used the 15 high-quality sequences of 22 kb region on *D. melanogaster* chromosome X from Sanger sequencing as “gold standard” and looked at potential errors and biases in DPGP V1 dataset by comparing the two. Note that though 15 Sanger sequences are from California population and the 37 Illumina/Solexa sequences from North Carolina, no significant divergence is expected based on a microsatellite survey of California and Florida populations (IRVIN *et al.* 1998).

Assuming neutrality across this 22kb region,  $S=314$  (100%) were expected. However,  $S=707$  (225%) were observed if missing information (N) is ignored at variant

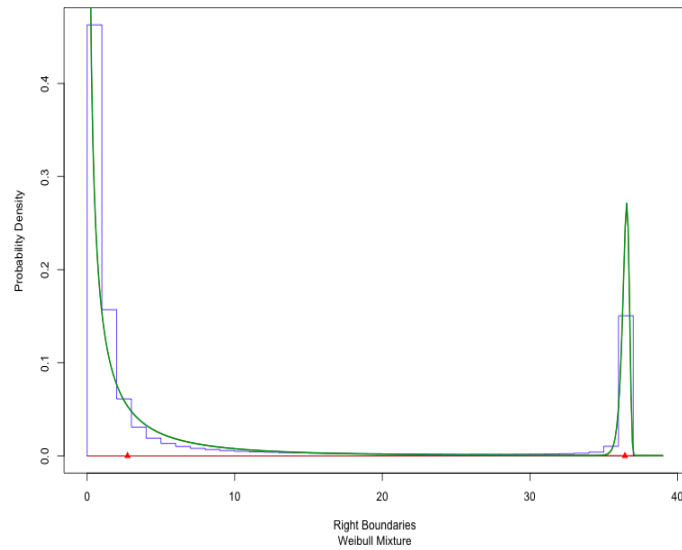
sites, and only  $S=228$  (72%) if all sites with at least one individual having N's are excluded. Using the former variants set with N's ignored will lead to many false SNP called by sequencing errors and assembly errors, while using the latter SNP set for downstream analysis would greatly compromise power due to the fact that about 30% of variants might be missing for analysis. In total, 10%~12% of all bases sequenced and called are classified to be missing (N's). By comparing the DPGP V1 polymorphic sites, it is obvious that huge proportion of SNP sites contain missing base calls for one or more individuals (Fig. A.1).



A



B



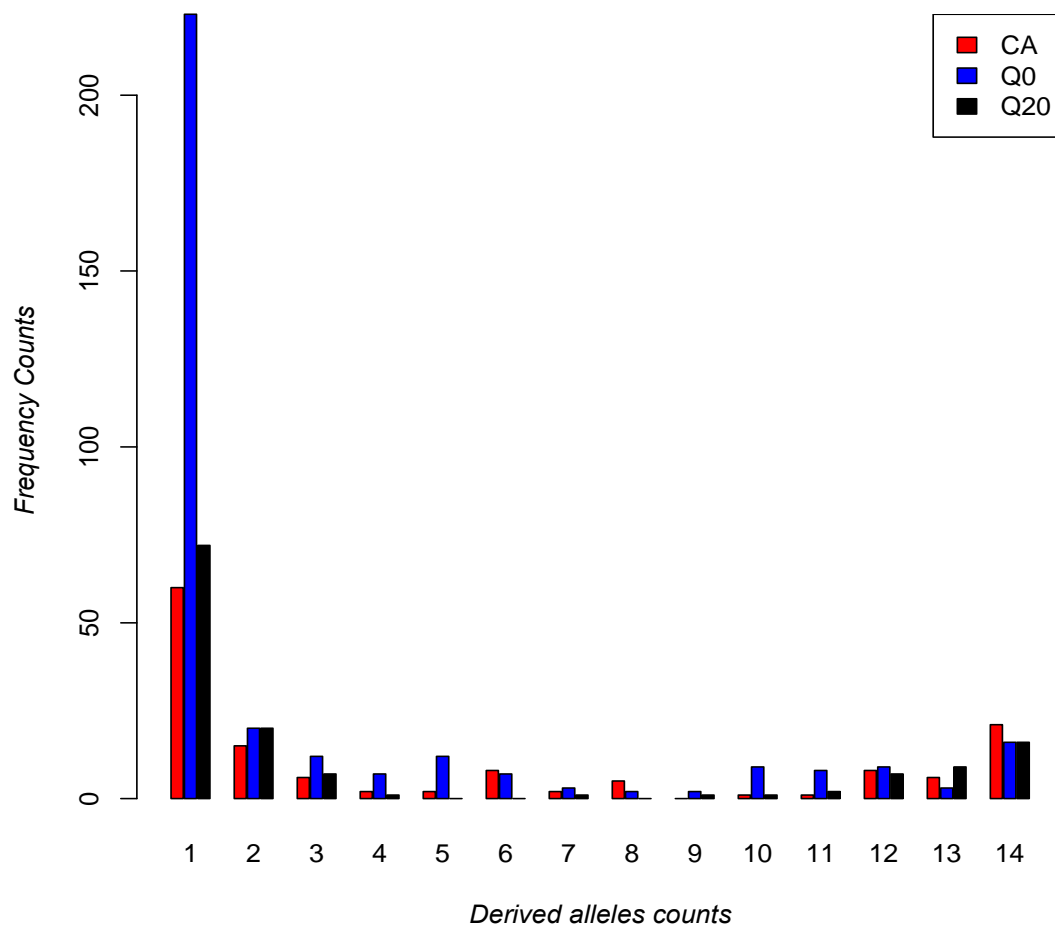
**Figure A.2: Spatial pattern of missing data distribution of a 2 Mb window on X chromosome.**

**A. Heterogeneity in sample size (number of individuals called to be N's at a nucleotide site) along the X chromosome. B. The frequency at which the number of individuals missing across sites. Only sites with at least one N are included. Observed (black bars), Weibull fitting curve (green).**

We then looked at how spatially the missing sites are distributed by randomly selecting a 2 Mb region from X chromosome (coordinates: 2,684,000–4,684,000). It is found that the spatial distribution of missing base calls is very heterogeneous, and can be approximately modeled using a mixture of Weibull distributions (Fig. A.2). This allows us to capture the spatial pattern (both amount and distribution) of missing information present in next-gen sequencing and with such pattern built into simulation studies, people can simulate sequences that mimic the true low-quality next-generation sequencing datasets for various purposes.

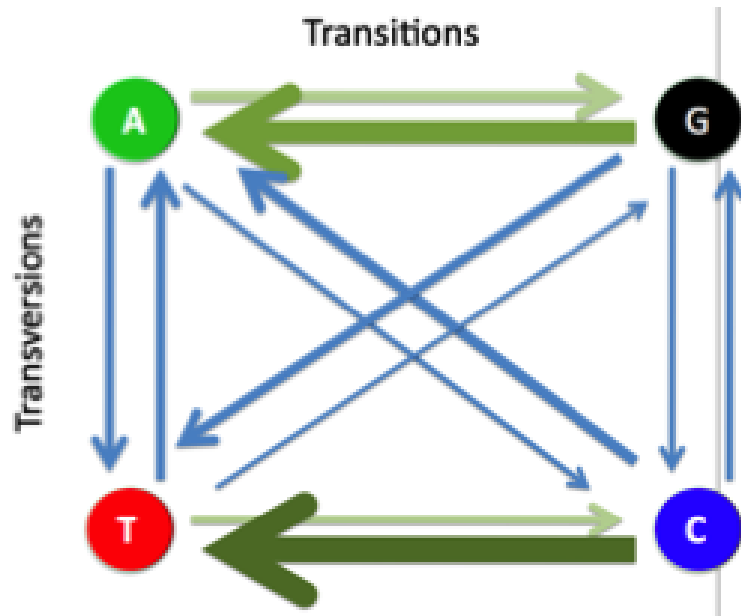
Solexa/Illumina sequences also showed an excess of rare alleles, especially singletons and doubletons compared with Sanger sequences, which are likely to be sequence errors (Fig. A.3). Adopting a stringent quality cutoff, most of singleton errors are filtered out though some errors are still kept exhibiting a constant excess of singletons. Interestingly, such quality filtering did not show observable quality improvement over the doubletons call set and there was still deficiency in high frequency derived alleles being called. To further investigate what error types mainly exist in the erroneous singleton calls, we compared the mismatches between the low-quality singleton calls (singletons have Phred quality scores of less than 20) and their ancestral states. The *D. melanogaster* genome showed substantial transitions from G to A and from C to T (SINGH *et al.* 2005), while such erroneous singleton calls tend to have a much more even rate of transitions and transversions with greatly elevated level of

sequencing/calling transversions (Fig. A.4B). One possible explanation of this bias towards sequencing/calling transversions is the result of simple sequencing bias that Solexa/Illumina first generation sequencer tends to have lower quality for some nucleotides than others. But this is not the case in DPGP V1 dataset: by calculating the proportion of bases with  $Q < 20$  among for each nucleotide using the entire X chromosome, there is no noticeable difference in nucleotide-specific sequencing quality overall, thus suggesting the sequencing technique employed did not discriminate any specific nucleotide base in assigning quality (Fig. A.5).



**Figure A.3: Polarized site-frequency spectrum (SFS) of 22kb on chromosome X comparing Sanger sequences (red), raw Solexa sequences (blue) and Solexa sequences with  $Q \geq 20$  filtering (black).**

A



B

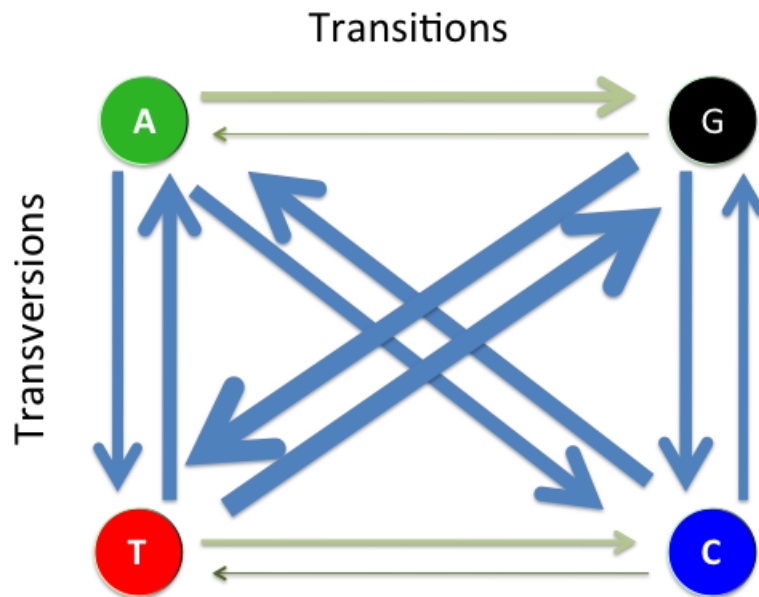
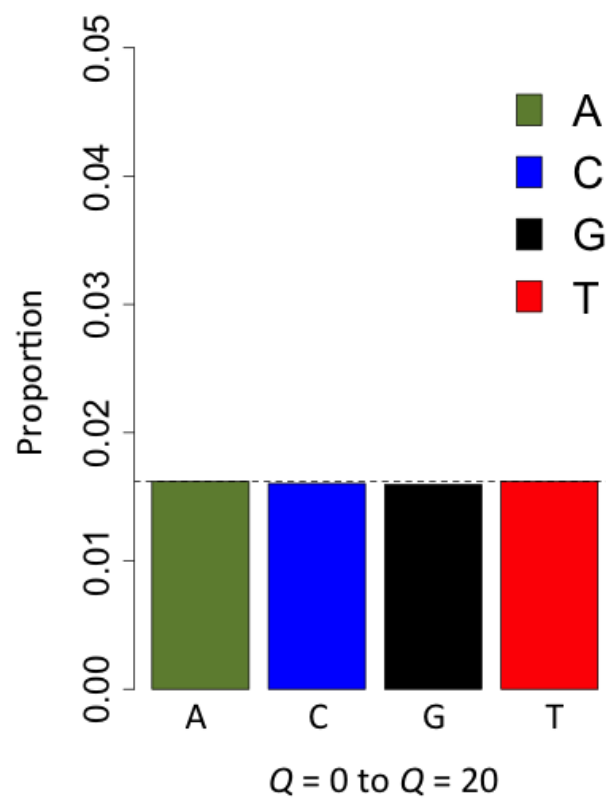


Figure A.4: DPGP V1.0 sequences contain substantial transversion errors and fewer transition errors.

A. The *D. melanogaster* genome average rate of transitions and transversions (SINGH *et al.* 2005). B. The transitions and transversions observed by comparing the raw Solexa/Illumina sequences with their ancestral states for the X chromosome (~22 Mb).



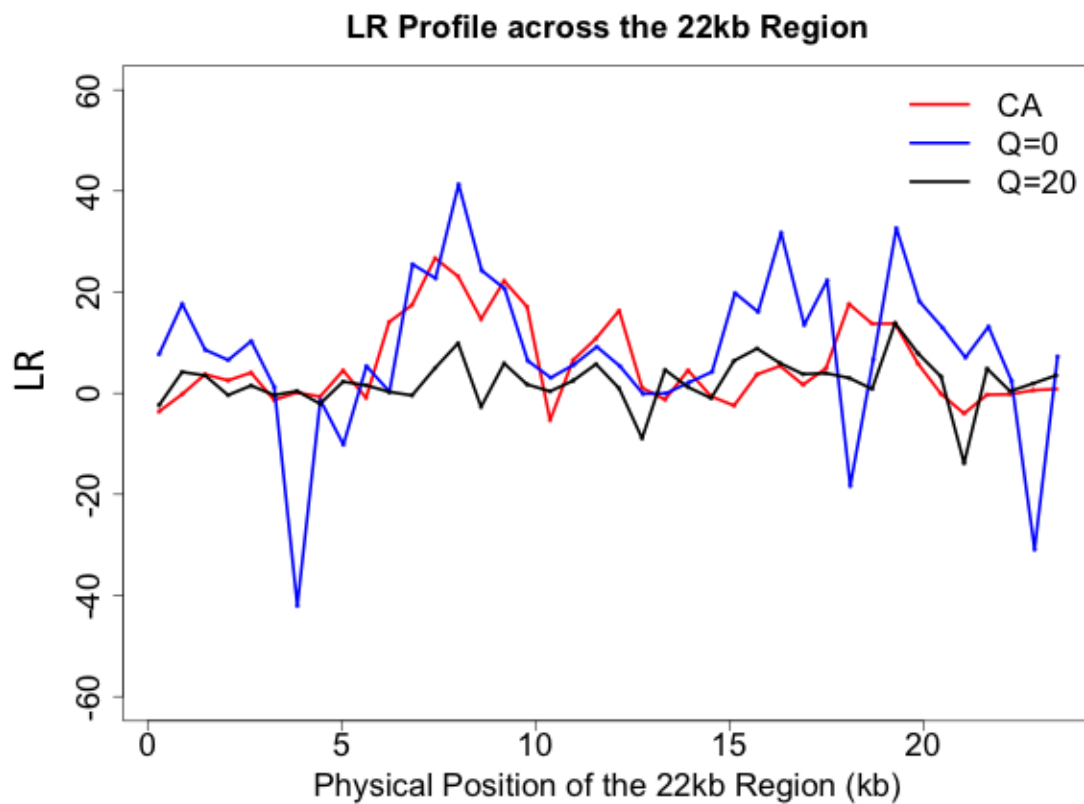


**Figure A.5: DPGP V1.0 has no bias toward sequencing a particular type of nucleotide acid with low quality.**

For each nucleotide, the proportion is calculated as the number of all of its base calls with  $Q < 20$  in the raw Solexa/Illumina sequences to the total number of the nucleotide base calls for X chromosome across all 37 individuals. Horizontal black dashed line shows the level of Adenine.

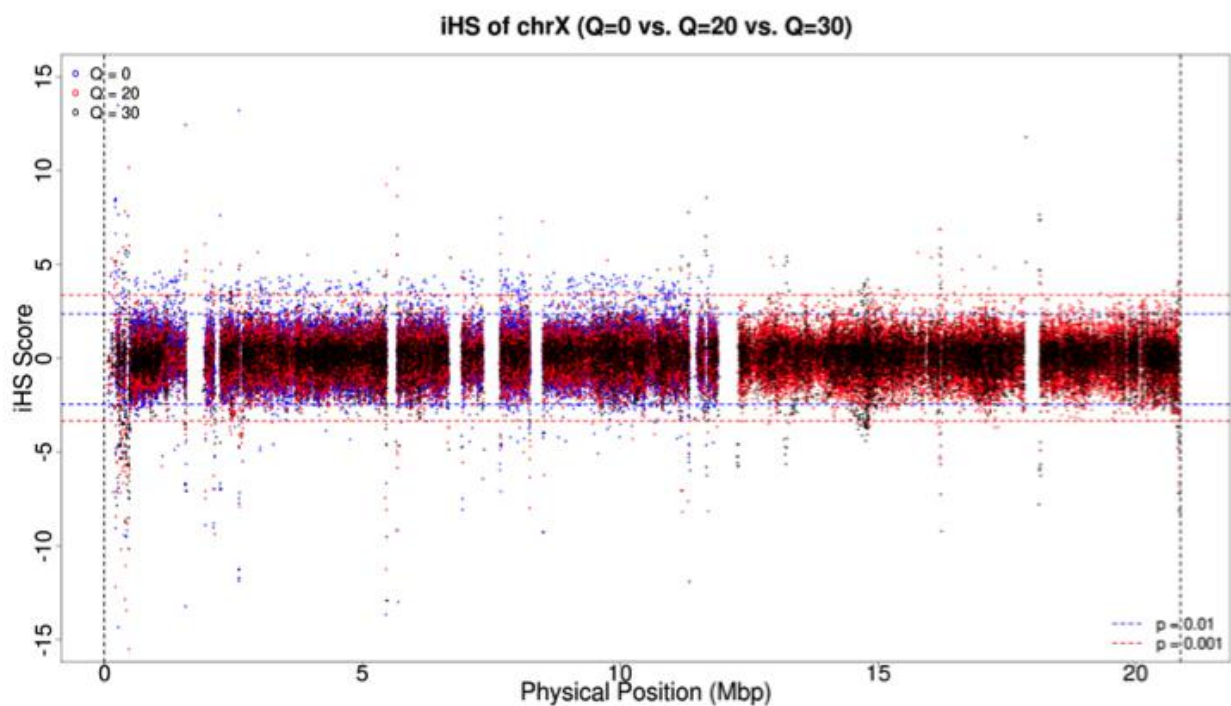
Sequence errors and missing data could greatly compromise the performance of SweepFinder. SweepFinder was able to discover two significant regions (coordinates within the 22 kb region: 7~9 kb, ~18 kb) to be targets of recent putative sweeps in the 22 kb segment around the notch gene if applied on the high-quality Sanger dataset of only 15 individuals (DUMONT and AQUADRO 2005). However, when analyzed with the raw DPGP V1 sequences, which contain substantial errors and biases, the second peak was not identified under the very noisy likelihood ratio (LR) pattern across the entire segment. With the extent of missing data dramatically increased after filtering with  $Q \geq 20$ , analysis using DPGP V1 sequences became very insensitive and failed to show any footprints of recent selective sweeps (Fig. A.6).

Since it is reasonably robust to sequence errors and missing data (Chapter 2), we applied iHS to identify putative targets of recent sweeps (Fig. A.7). The number of SNPs of chromosome X across all 37 lines if excluding missing information and gaps, decreases from 396,128 (raw release sequences,  $Q \geq 0$ ), 160,064 ( $Q \geq 20$ ), to 40,746 ( $Q \geq 30$ ), and the number of iHS scores calculated from about 68,000 (raw release sequences,  $Q \geq 0$ ), 65,700 ( $Q \geq 20$ ) to 21,500 ( $Q \geq 30$ ).



**Figure A.6: Performance of SweepFinder on datasets with varying level of quality and missing data.**

SweepFinder was applied on the 15-individual sample (CA, red), raw DPGP V1 sequences (blue) and DPGP V1 sequences after  $Q \geq 20$  filtering. Note the locus at about coordinate 18 kb showed significance with CA dataset where both DPGP V1 datasets had insignificant values.



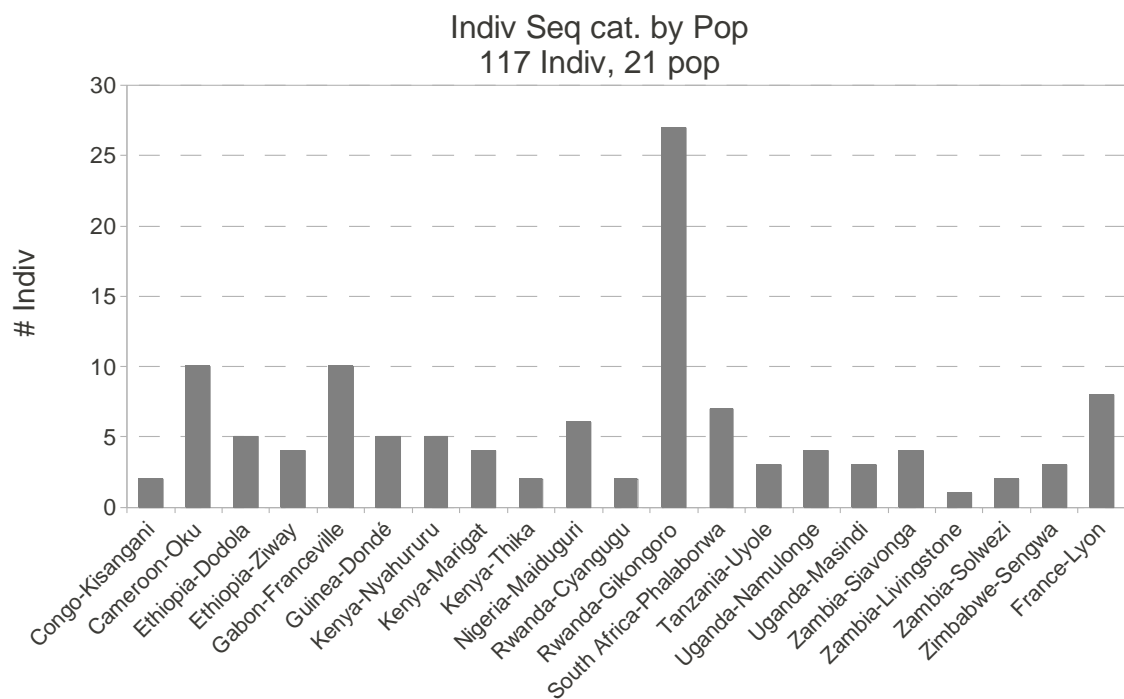
**Figure A.7: Identification of putative targets of recent selective sweeps by iHS using DPGP V1 datasets with varying quality filtering.**

The horizontal blue dashed lines are confidence interval for FDR=5% while the red dashed lines are that for FDR=1%.

### A.3.2 Sequencing errors and biases in DPGP V2

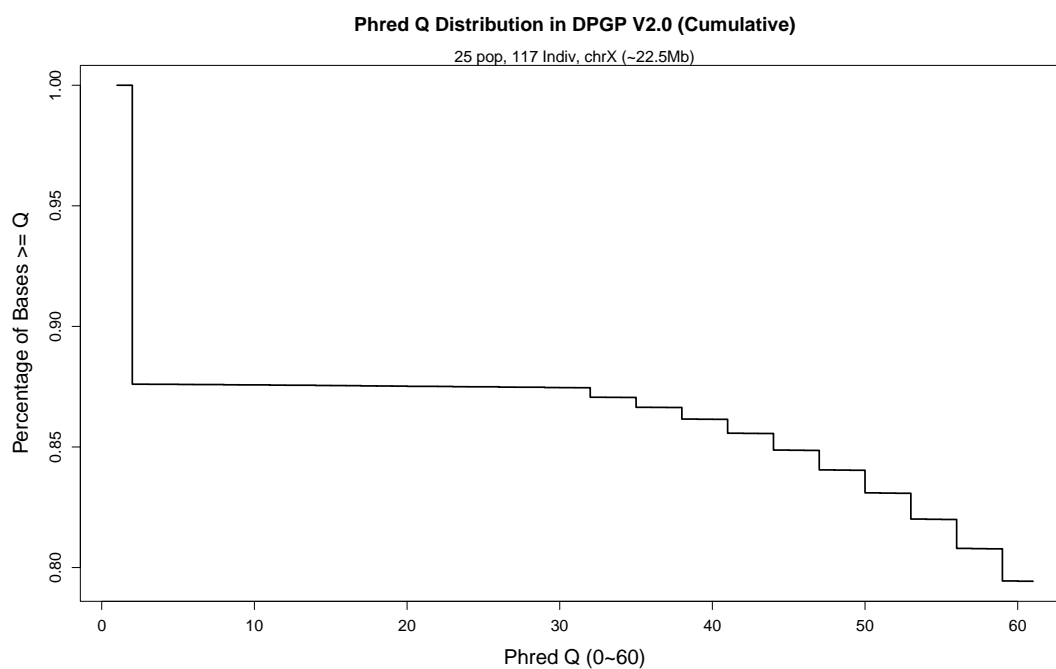
The 117 individuals were sampled from 21 different populations in Africa, and samples from Gikongoro, Rwanda population have the largest sample size of 27 individuals (Fig. A.8). The Rwanda population is thought to be an ancestral group and no signs of IBD was observed in DPGP V2. For all 117 individuals, there are about 12.7% sites missing/masked with N's and most of the bases called have very high quality ( $Q \geq 30$ ) (Fig. A.9A). We also observed very similar pattern for the Rwanda population, with slightly more sites being missing (Fig. A.9B).

We selected the same 2 Mb segment based on FlyBase Release v5.3 from the Rwanda population for further analysis (chrX: 2,684,000~4,684,000 bp). There is substantial proportion of variant sites that are affected by missing base calls, and 51% of the variant sites contain at least one missing base calls across the 27 individuals (Fig. A.10A). Individually, each sequence had 12%~15% of both its variant and invariant sites not called and the rest of 85%~88% of bases in this 2 Mbp region that were successfully called generally have very high quality. Surprisingly, there is only slight decrease in bases retained even after Phred quality filtering of  $Q \geq 40$  (Fig. A.10B). Such bipolar pattern of either very high quality base calls or none (N's) could be the result of extensive modeling and correction of the dataset by the DPGP panel.

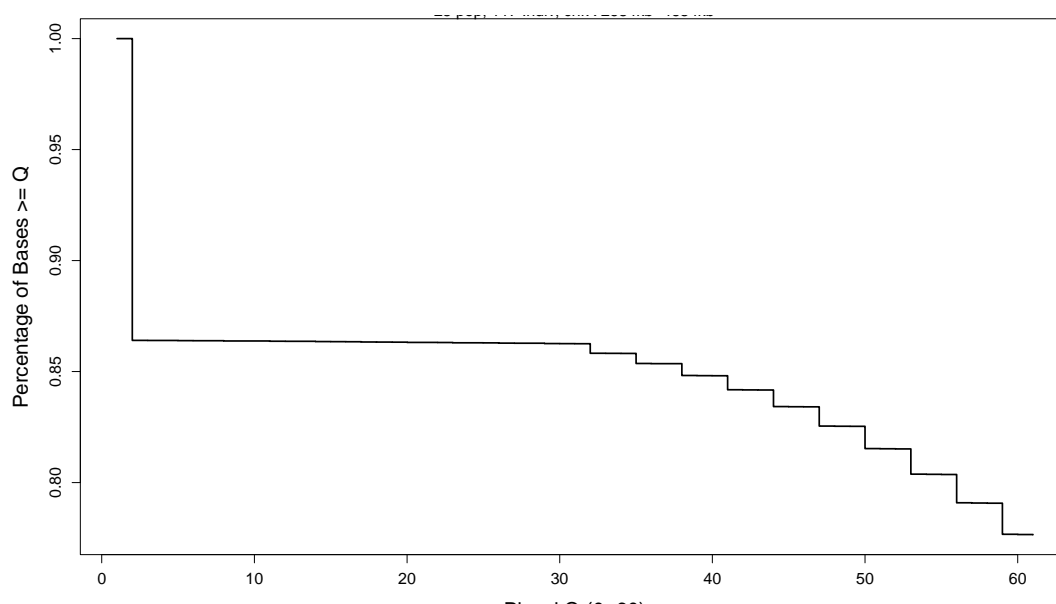


**Figure A.8: Number of *D. melanogaster* individuals sequenced in each population in DPGP V2.**

**A**



**B**



**Figure A.9: Base call qualities of DPGP V2 sequences.**

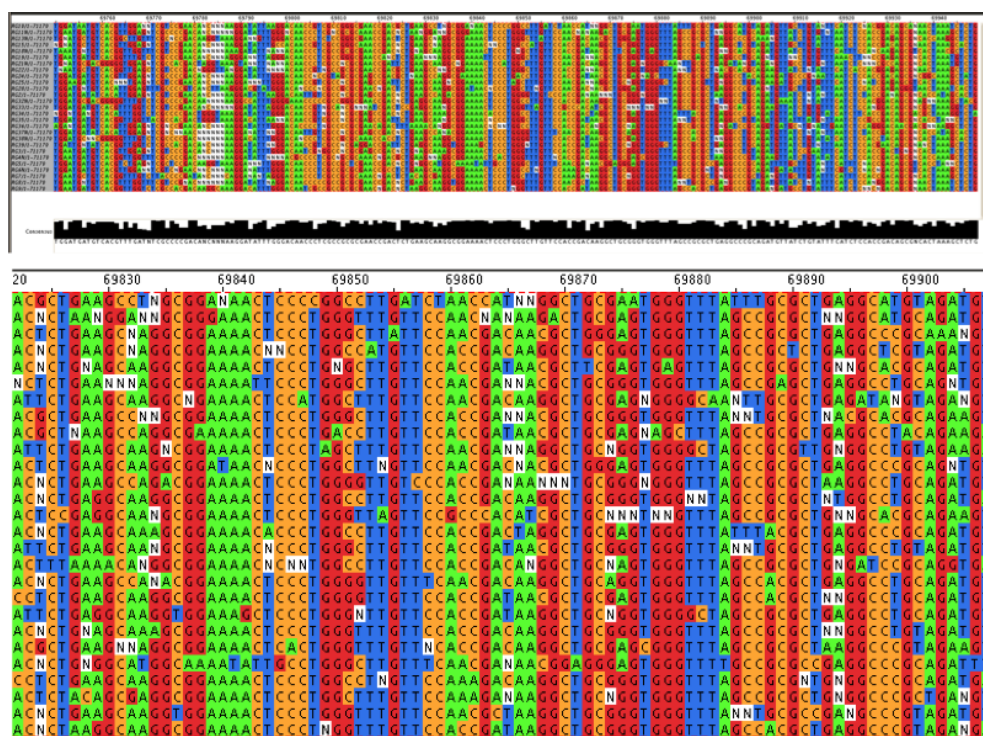
**A. All 117 individuals in DPGP V2. B. The 27 individuals of Rwanda population.**

**Figure A.10: Analysis of a 2 Mb region in RG population with all 27 individuals.**

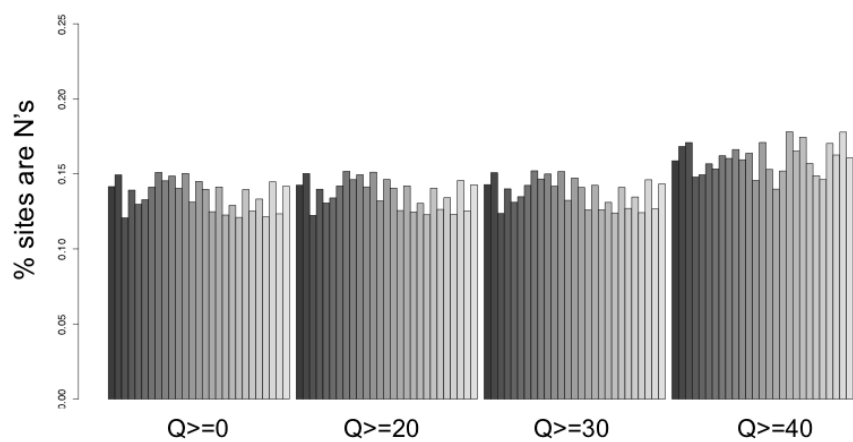
**A. Only polymorphic sites are included. The top panel is an overview of the SNP sites. The middle panel black bars' height shows the variation in sample size at each SNP site. And the bottom panel is zoom-in of the SNP sites. Missing sites are labeled with N's in white boxes. B. The proportion of nucleotide sites (both variant or invariant in the population) that are filtered out under varying Phred quality cutoffs for each individual.**



A

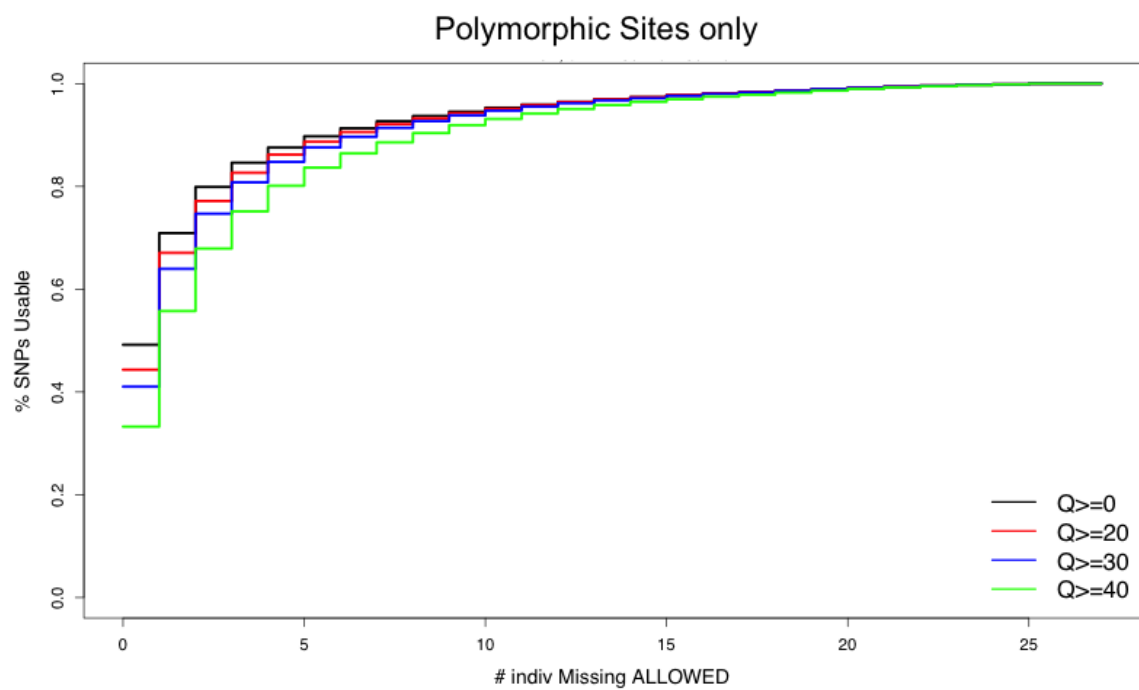


B

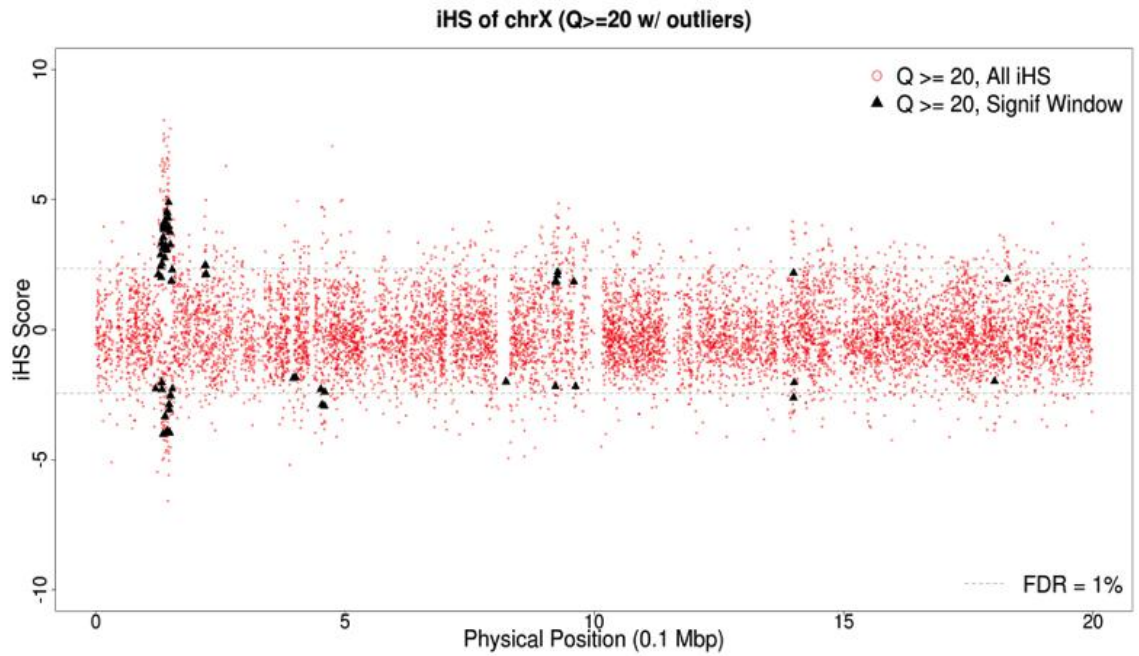


We also investigated the proportion of variant sites that are available for analysis under different combinations of base quality filtering and sample size requirements for varying research purposes. Due to the extensive correction by the DPGP panel, quality filtering results in only slight difference in available SNPs when sample completeness requirement is very high (less than 4 missing out of 27 individuals) and negligible effects when sample size requirement is relatively flexible (Fig. A.11). In this study, data completeness is a severe concern for following analysis and is insensitive to quality filtering.

iHS was then applied on this 2-Mb segment to detect putative targets of recent sweeps (Fig. A.12). There are 53 windows that are evaluated to be statistically significant and region chrX: 2,820,000~2,870,000 bp demonstrated very strong signals of recent selective sweeps.



**Figure A.11: Proportion of variant sites that can be used under varying quality filtering and sample size requirements.**



**Figure A.12: Putative targets identified by iHS for the X chromosome in Rwanda population.**

Raw DPGP V2 sequences were filtered first with  $Q \geq 20$ . The iHS scores at variant sites are labeled in red dots and significant iHS sliding windows are labeled with black triangles under FDR=1%. The 99% confidence interval of single iHS score is labeled with the two black horizontal dashed lines.

## **A.4 DISCUSSION**

Relatively low-pass sequencing strategy is commonly used in many species, which might impose computational challenges for downstream bioinformatics analysis. Certain population studies such as genome wide association studies (GWAS) benefit greatly from low-depth sequencing of a large number of individuals coupling with compensating techniques such as imputation (PORCU *et al.* 2013). While for other research goals, people need to evaluate the potential benefits and the detrimental effects as Chapter 2 pointed out.

We analyzed both DPGP V1 and V2 release of *Drosophila melanogaster* samples and found that there are substantial sequencing errors and biases in DPGP V1.0 release of 37 North American lines. Data quality is overall optimistic in DPGP V2 and most of bases called have very high quality ( $Q \geq 30$ ). However, for both datasets, there are massive amount of missing information after quality filtering, which is a great for population genetic inferences (JENSEN *et al.* 2008b; POOL *et al.* 2010). We confirmed that singletons and doubletons were the primary sources of sequencing errors and the SFS is negatively skewed even after quality filtering in DPGP V1. There is no significant correlation between error rate and the type of nucleotides, but the spatial pattern of how such low-quality bases and missing data are distributed in a very heterogeneous way.

The sequencing errors and biases may be substantially alleviated by employing longer reads (SHARON *et al.* 2013) and carefully preparing the libraries and properly designing the experiments according to the research goals (WANG *et al.* 2009; PAREEK *et al.* 2011; TARIQ *et al.* 2011; VAN DIJK *et al.* 2014). Ultra-deep sequencing can be an alternative too, but is not feasible for most projects due to budget constraints (LIGHTEN *et al.* 2014). Targeted deep sequencing seems to be another option but target enrichment adds another layer of potential biases and errors (MAMANOVA *et al.* 2010; LEPROUST 2012). Single-cell sequencing is a very promising technique that allows read length of several kb. However, its accuracy, commercial applications as well as costs are yet to be fully developed (KORFHAGE *et al.* 2013; SHAPIRO *et al.* 2013). In addition, there are still needs for more sophisticated software designed specifically for low-quality/low-depth datasets for various purposes that will help out accurately make more useful inferences from the constrained datasets.

## References

- AGARWAL, A., D. KOPPSTEIN, J. ROZOWSKY, A. SBONER, L. HABEGGER *et al.*, 2010 Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics* **11**: 383.
- AHN, S. M., T. H. KIM, S. LEE, D. KIM, H. GHANG *et al.*, 2009 The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19**: 1622-1629.
- AIRD, D., M. G. ROSS, W. S. CHEN, M. DANIELSSON, T. FENNELLS *et al.*, 2011 Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**: R18.
- AKEY, J. M., G. ZHANG, K. ZHANG, L. JIN and M. D. SHRIVER, 2002 Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12**: 1805-1814.
- ALACHOTIS, N., A. STAMATAKIS and P. PAVLIDIS, 2012 OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *Bioinformatics* **28**: 2274-2275.
- ANDOLFATTO, P., 2007 Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res* **17**: 1755-1762.
- ANDOLFATTO, P., and M. PRZEWORSKI, 2000 A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* **156**: 257-268.
- AUER, P. L., and R. W. DOERGE, 2010 Statistical design and analysis of RNA sequencing data. *Genetics* **185**: 405-416.
- BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**: 519-520.
- BENTLEY, D. R., S. BALASUBRAMANIAN, H. P. SWERDLOW, G. P. SMITH, J. MILTON *et al.*, 2008 Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.
- BHANGALE, T. R., M. J. RIEDER and D. A. NICKERSON, 2008 Estimating coverage and power for genetic association studies using near-complete variation data. *Nat Genet* **40**: 841-843.
- BRADNAM, K. R., J. N. FASS, A. ALEXANDROV, P. BARANAY, M. BECHNER *et al.*, 2013 Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**: 10.
- BULLARD, J. H., E. PURDOM, K. D. HANSEN and S. DUDOIT, 2010 Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**: 94.
- BUSTAMANTE, C. D., R. NIELSEN, S. A. SAWYER, K. M. OLSEN, M. D. PURUGGANAN *et al.*, 2002 The cost of inbreeding in *Arabidopsis*. *Nature* **416**: 531-534.
- CARLSON, C. S., D. J. THOMAS, M. A. EBERLE, J. E. SWANSON, R. J. LIVINGSTON *et al.*, 2005 Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res* **15**: 1553-1565.

- CHARLESWORTH, J., and A. EYRE-WALKER, 2008 The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol* **25**: 1007-1015.
- CLARK, A. G., S. GLANOWSKI, R. NIELSEN, P. D. THOMAS, A. KEJARIWAL *et al.*, 2003 Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**: 1960-1963.
- CRAWFORD, J. E., and B. P. LAZZARO, 2012 Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Front Genet* **3**: 66.
- CRISCI, J. L., Y. P. POH, S. MAHAJAN and J. D. JENSEN, 2013 The impact of equilibrium assumptions on tests of selection. *Front Genet* **4**: 235.
- DAINES, B., H. WANG, Y. LI, Y. HAN, R. GIBBS *et al.*, 2009 High-throughput multiplex sequencing to discover copy number variants in *Drosophila*. *Genetics* **182**: 935-941.
- DEPRISTO, M. A., E. BANKS, R. POPLIN, K. V. GARIMELLA, J. R. MAGUIRE *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491-498.
- DOHM, J. C., C. LOTTAZ, T. BORODINA and H. HIMMELBAUER, 2008 Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**: e105.
- DONIGER, S. W., H. S. KIM, D. SWAIN, D. CORCUERA, M. WILLIAMS *et al.*, 2008 A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet* **4**: e1000183.
- DUMONT, V. B., and C. F. AQUADRO, 2005 Multiple signatures of positive selection downstream of notch on the X chromosome in *Drosophila melanogaster*. *Genetics* **171**: 639-653.
- EARL, D., K. BRADNAM, J. ST JOHN, A. DARLING, D. LIN *et al.*, 2011 Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* **21**: 2224-2241.
- EMDE, A. K., M. H. SCHULZ, D. WEESE, R. SUN, M. VINGRON *et al.*, 2012 Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. *Bioinformatics* **28**: 619-627.
- ERSOZ, E. S., M. H. WRIGHT, J. L. PANGILINAN, M. J. SHEEHAN, C. TOBIAS *et al.*, 2012 SNP discovery with EST and NextGen sequencing in switchgrass (*Panicum virgatum* L.). *PLoS One* **7**: e44112.
- EWING, B., and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**: 186-194.
- EWING, B., L. HILLIER, M. C. WENDL and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175-185.
- FISTON-LAVIER, A. S., N. D. SINGH, M. LIPATOV and D. A. PETROV, 2010 *Drosophila melanogaster* recombination rate calculator. *Gene* **463**: 18-20.
- GALTIER, N., F. DEPAULIS and N. H. BARTON, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* **155**: 981-987.



- GENOMES PROJECT, C., G. R. ABECASIS, D. ALTSHULER, A. AUTON, L. D. BROOKS *et al.*, 2010 A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.
- GENOMES PROJECT, C., G. R. ABECASIS, A. AUTON, L. D. BROOKS, M. A. DEPRISTO *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56-65.
- GLINKA, S., L. OMETTO, S. MOUSSET, W. STEPHAN and D. DE LORENZO, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* **165**: 1269-1278.
- GRAVELEY, B. R., A. N. BROOKS, J. W. CARLSON, M. O. DUFF, J. M. LANDOLIN *et al.*, 2011 The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**: 473-479.
- GROSSMAN, S. R., I. SHLYAKHTER, E. K. KARLSSON, E. H. BYRNE, S. MORALES *et al.*, 2010 A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**: 883-886.
- GUO, Y., J. LI, C. I. LI, J. LONG, D. C. SAMUELS *et al.*, 2012 The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13**: 666.
- GUTTMAN, M., M. GARBER, J. Z. LEVIN, J. DONAGHEY, J. ROBINSON *et al.*, 2010 Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**: 503-510.
- HABEGGER, L., A. SBONER, T. A. GIANOULIS, J. ROZOWSKY, A. AGARWAL *et al.*, 2011 RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27**: 281-283.
- HANSEN, K. D., S. E. BRENNER and S. DUDOIT, 2010 Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**: e131.
- HELLMANN, I., Y. MANG, Z. GU, P. LI, F. M. DE LA VEGA *et al.*, 2008 Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res* **18**: 1020-1029.
- HELYAR, S. J., M. T. LIMBORG, D. BEKKEVOLD, M. BABBUCCI, J. VAN HOUDT *et al.*, 2012 SNP discovery using Next Generation Transcriptomic Sequencing in Atlantic herring (*Clupea harengus*). *PLoS One* **7**: e42089.
- HERNANDEZ, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* **24**: 2786-2787.
- HILLIER, L. W., V. REINKE, P. GREEN, M. HIRST, M. A. MARRA *et al.*, 2009 Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. *Genome Res* **19**: 657-666.
- HOLLOWAY, A. K., D. J. BEGUN, A. SIEPEL and K. S. POLLARD, 2008 Accelerated sequence divergence of conserved genomic elements in *Drosophila melanogaster*. *Genome Res* **18**: 1592-1601.
- HUDSON, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337-338.
- HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153-159.

- HUELSENBECK, J. P., and F. RONQUIST, 2001 MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754-755.
- HUNKAPILLER, T., R. J. KAISER, B. F. KOOP and L. HOOD, 1991 Large-scale and automated DNA sequence determination. *Science* **254**: 59-67.
- HUTCHISON, C. A., 3RD, 2007 DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res* **35**: 6227-6237.
- JENSEN, J. D., V. L. BAUER DUMONT, A. B. ASHMORE, A. GUTIERREZ and C. F. AQUADRO, 2007a Patterns of sequence variability and divergence at the diminutive gene region of *Drosophila melanogaster*: complex patterns suggest an ancestral selective sweep. *Genetics* **177**: 1071-1085.
- JENSEN, J. D., Y. KIM, V. B. DUMONT, C. F. AQUADRO and C. D. BUSTAMANTE, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**: 1401-1410.
- JENSEN, J. D., K. R. THORNTON and P. ANDOLFATTO, 2008a An approximate bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genet* **4**: e1000198.
- JENSEN, J. D., K. R. THORNTON and C. F. AQUADRO, 2008b Inferring selection in partially sequenced regions. *Mol Biol Evol* **25**: 438-446.
- JENSEN, J. D., K. R. THORNTON, C. D. BUSTAMANTE and C. F. AQUADRO, 2007b On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics* **176**: 2371-2379.
- JIANG, R., S. TAVARE and P. MARJORAM, 2009 Population genetic inference from resequencing data. *Genetics* **181**: 187-197.
- JOHNSON, P. L., and M. SLATKIN, 2006 Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res* **16**: 1320-1327.
- JOHNSON, P. L., and M. SLATKIN, 2008 Accounting for bias from sequencing error in population genetic estimates. *Mol Biol Evol* **25**: 199-206.
- JUAREZ, M. T., R. A. PATTERSON, E. SANDOVAL-GUILLEN and W. MCGINNIS, 2011 Duox, Flotillin-2, and Src42A are required to activate or delimit the spread of the transcriptional response to epidermal wounds in *Drosophila*. *PLoS Genet* **7**: e1002424.
- KAMPA, D., J. CHENG, P. KAPRANOV, M. YAMANAKA, S. BRUBAKER *et al.*, 2004 Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* **14**: 331-342.
- KATANAIEV, V. L., G. P. SOLIS, G. HAUSMANN, S. BUESTORF, N. KATANAYEVA *et al.*, 2008 Reggie-1/flotillin-2 promotes secretion of the long-range signalling forms of Wingless and Hedgehog in *Drosophila*. *EMBO J* **27**: 509-521.
- KEIGHTLEY, P. D., and A. EYRE-WALKER, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**: 2251-2261.
- KEIGHTLEY, P. D., U. TRIVEDI, M. THOMSON, F. OLIVER, S. KUMAR *et al.*, 2009 Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res* **19**: 1195-1201.

- KELLEY, J. L., and W. J. SWANSON, 2008 Positive selection in the human genome: from genome scans to biological significance. *Annu Rev Genomics Hum Genet* **9**: 143-160.
- KIM, S. Y., K. E. LOHMUELLER, A. ALBRECHTSEN, Y. LI, T. KORNELIUSSEN *et al.*, 2011 Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* **12**: 231.
- KIM, Y., and R. NIELSEN, 2004 Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**: 1513-1524.
- KIM, Y., and W. STEPHAN, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**: 765-777.
- KOBOLDT, D. C., K. CHEN, T. WYLIE, D. E. LARSON, M. D. MCLELLAN *et al.*, 2009 VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**: 2283-2285.
- KOBOLDT, D. C., K. M. STEINBERG, D. E. LARSON, R. K. WILSON and E. R. MARDIS, 2013 The next-generation sequencing revolution and its impact on genomics. *Cell* **155**: 27-38.
- KORFHAGE, C., E. FISCH, E. FRICKE, S. BAEDKER and D. LOEFFERT, 2013 Whole-genome amplification of single-cell genomes for next-generation sequencing. *Curr Protoc Mol Biol* **104**: 7 14 11-17 14 11.
- KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**: 412-417.
- LANGLEY, C. H., K. STEVENS, C. CARDENO, Y. C. LEE, D. R. SCHRIDER *et al.*, 2012 Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* **192**: 533-598.
- LANGMEAD, B., C. TRAPNELL, M. POP and S. L. SALZBERG, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- LEPROUST, E., 2012 Target enrichment strategies for next generation sequencing. *MLO Med Lab Obs* **44**: 26-27.
- LEVIN, J. Z., M. YASSOUR, X. ADICONIS, C. NUSBAUM, D. A. THOMPSON *et al.*, 2010 Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**: 709-715.
- LI, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987-2993.
- LI, H., and R. DURBIN, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.
- LI, H., B. HANDSAKER, A. WYSOKER, T. FENNELL, J. RUAN *et al.*, 2009a The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.
- LI, H., J. RUAN and R. DURBIN, 2008a Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**: 1851-1858.
- LI, H., and W. STEPHAN, 2005 Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics* **171**: 377-384.

- LI, H., and W. STEPHAN, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet* **2**: e166.
- LI, J., H. JIANG and W. H. WONG, 2010a Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol* **11**: R50.
- LI, R., Y. LI, K. KRISTIANSEN and J. WANG, 2008b SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**: 713-714.
- LI, R., C. YU, Y. LI, T. W. LAM, S. M. YIU *et al.*, 2009b SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**: 1966-1967.
- LI, Y., S. L. YANG, Z. L. TANG, W. T. CUI, Y. L. MU *et al.*, 2010b Expression and SNP association analysis of porcine FBXL4 gene. *Mol Biol Rep* **37**: 579-585.
- LIGHTEN, J., C. VAN OOSTERHOUT, I. G. PATERSON, M. McMULLAN and P. BENTZEN, 2014 Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Mol Ecol Resour*.
- LIU, C. M., T. WONG, E. WU, R. LUO, S. M. YIU *et al.*, 2012 SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics* **28**: 878-879.
- LYNCH, M., 2008 Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol Biol Evol* **25**: 2409-2419.
- LYNCH, M., 2009 Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* **182**: 295-301.
- MACKAY, T. F., S. RICHARDS, E. A. STONE, A. BARBADILLA, J. F. AYROLES *et al.*, 2012 The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**: 173-178.
- MAMANOVA, L., A. J. COFFEY, C. E. SCOTT, I. KOZAREWA, E. H. TURNER *et al.*, 2010 Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7**: 111-118.
- MARCHINI, J., and B. HOWIE, 2010 Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**: 499-511.
- MARDIS, E. R., 2008 Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**: 387-402.
- MARGUERAT, S., and J. BAHLER, 2010 RNA-seq: from technology to biology. *Cell Mol Life Sci* **67**: 569-579.
- MARGULIES, M., M. EGHOLM, W. E. ALTMAN, S. ATTIYA, J. S. BADER *et al.*, 2005 Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.
- MARIONI, J. C., M. WHITE, S. TAVARE and A. G. LYNCH, 2008 Hidden copy number variation in the HapMap population. *Proc Natl Acad Sci U S A* **105**: 10067-10072.
- MARYGOLD, S. J., P. C. LEYLAND, R. L. SEAL, J. L. GOODMAN, J. THURMOND *et al.*, 2013 FlyBase: improvements to the bibliography. *Nucleic Acids Res* **41**: D751-757.
- McKENNA, A., M. HANNA, E. BANKS, A. SIVACHENKO, K. CIBULSKIS *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

- MORTAZAVI, A., B. A. WILLIAMS, K. MCCUE, L. SCHAEFFER and B. WOLD, 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621-628.
- NAGALAKSHMI, U., K. WAERN and M. SNYDER, 2010 RNA-Seq: a method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol* **Chapter 4**: Unit 4 11 11-13.
- NECHAEV, S., D. C. FARGO, G. DOS SANTOS, L. LIU, Y. GAO *et al.*, 2010 Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327**: 335-338.
- NI, T., D. L. CORCORAN, E. A. RACH, S. SONG, E. P. SPANA *et al.*, 2010 A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* **7**: 521-527.
- NIELSEN, R., C. BUSTAMANTE, A. G. CLARK, S. GLANOWSKI, T. B. SACKTON *et al.*, 2005a A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.
- NIELSEN, R., I. HELLMANN, M. HUBISZ, C. BUSTAMANTE and A. G. CLARK, 2007 Recent and ongoing selection in the human genome. *Nat Rev Genet* **8**: 857-868.
- NIELSEN, R., M. J. HUBISZ, I. HELLMANN, D. TORGERSON, A. M. ANDRES *et al.*, 2009 Darwinian and demographic forces affecting human protein coding genes. *Genome Res* **19**: 838-849.
- NIELSEN, R., S. WILLIAMSON, Y. KIM, M. J. HUBISZ, A. G. CLARK *et al.*, 2005b Genomic scans for selective sweeps using SNP data. *Genome Res* **15**: 1566-1575.
- OLEKSYK, T. K., K. ZHAO, F. M. DE LA VEGA, D. A. GILBERT, S. J. O'BRIEN *et al.*, 2008 Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS One* **3**: e1712.
- OMETTO, L., S. GLINKA, D. DE LORENZO and W. STEPHAN, 2005 Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Mol Biol Evol* **22**: 2119-2130.
- ONMUS-LEONE, F., J. HANG, R. J. CLIFFORD, Y. YANG, M. C. RILEY *et al.*, 2013 Enhanced de novo assembly of high throughput pyrosequencing data using whole genome mapping. *PLoS One* **8**: e61762.
- OSSOWSKI, S., K. SCHNEEBERGER, R. M. CLARK, C. LANZ, N. WARTHMAN *et al.*, 2008 Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* **18**: 2024-2033.
- OZSOLAK, F., and P. M. MILOS, 2011 RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**: 87-98.
- PAREEK, C. S., R. SMOCZYNSKI and A. TRETYN, 2011 Sequencing technologies and genome sequencing. *J Appl Genet* **52**: 413-435.
- PARK, P. J., 2009 ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**: 669-680.
- PAVLIDIS, P., J. D. JENSEN and W. STEPHAN, 2010 Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics* **185**: 907-922.

- PAVLIDIS, P., D. ZIVKOVIC, A. STAMATAKIS and N. ALACHIOTIS, 2013 SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol Biol Evol* **30**: 2224-2234.
- PEPKE, S., B. WOLD and A. MORTAZAVI, 2009 Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**: S22-32.
- POOL, J. E., and C. F. AQUADRO, 2006 History and structure of sub-Saharan populations of *Drosophila melanogaster*. *Genetics* **174**: 915-929.
- POOL, J. E., R. B. CORBETT-DETIG, R. P. SUGINO, K. A. STEVENS, C. M. CARDENO *et al.*, 2012 Population Genomics of sub-saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genet* **8**: e1003080.
- POOL, J. E., I. HELLMANN, J. D. JENSEN and R. NIELSEN, 2010 Population genetic inference from genomic sequence variation. *Genome Res* **20**: 291-300.
- PORCU, E., S. SANNA, C. FUCHSBERGER and L. G. FRITSCHKE, 2013 Genotype imputation in genome-wide association studies. *Curr Protoc Hum Genet* **Chapter 1**: Unit 1 25.
- PRESGRAVES, D. C., 2006 Intron length evolution in *Drosophila*. *Mol Biol Evol* **23**: 2203-2213.
- RIVERA, C. M., and B. REN, 2013 Mapping human epigenomes. *Cell* **155**: 39-55.
- ROBINSON, M. D., D. J. MCCARTHY and G. K. SMYTH, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.
- ROYCE, T. E., J. S. ROZOWSKY, P. BERTONE, M. SAMANTA, V. STOLC *et al.*, 2005 Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet* **21**: 466-475.
- SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. LEVINE, D. J. RICHTER *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832-837.
- SABETI, P. C., S. F. SCHAFFNER, B. FRY, J. LOHMUELLER, P. VARILLY *et al.*, 2006 Positive natural selection in the human lineage. *Science* **312**: 1614-1620.
- SABETI, P. C., P. VARILLY, B. FRY, J. LOHMUELLER, E. HOSTETTER *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913-918.
- SANGER, F., 1988 Sequences, sequences, and sequences. *Annu Rev Biochem* **57**: 1-28.
- SCHWARTZ, T. S., H. TAE, Y. YANG, K. MOCKAITIS, J. L. VAN HEMERT *et al.*, 2010 A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences. *BMC Genomics* **11**: 694.
- SHAPIRO, E., T. BIEZUNER and S. LINNARSSON, 2013 Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14**: 618-630.
- SHARON, D., H. TILGNER, F. GRUBERT and M. SNYDER, 2013 A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**: 1009-1014.

- SHEN, Y., Z. WAN, C. COARFA, R. DRABEK, L. CHEN *et al.*, 2010 A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res* **20**: 273-280.
- SHENDURE, J., R. D. MITRA, C. VARMA and G. M. CHURCH, 2004 Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5**: 335-344.
- SINGH, N. D., P. F. ARNDT and D. A. PETROV, 2005 Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. *Genetics* **169**: 709-722.
- SINGH, N. D., J. D. JENSEN, A. G. CLARK and C. F. AQUADRO, 2013 Inferences of demography and selection in an African population of *Drosophila melanogaster*. *Genetics* **193**: 215-228.
- SMITH, J. M., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. *Genet Res* **23**: 23-35.
- STAPLEY, J., J. REGER, P. G. FEULNER, C. SMADJA, J. GALINDO *et al.*, 2010 Adaptation genomics: the next generation. *Trends Ecol Evol* **25**: 705-712.
- STEPHAN, W., and H. LI, 2007 The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity (Edinb)* **98**: 65-68.
- STEPHAN, W., Y. S. SONG and C. H. LANGLEY, 2006 The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **172**: 2647-2663.
- STONE, E. A., 2012 Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines. *Genome Res* **22**: 966-974.
- SWERDLOW, H., and R. GESTELAND, 1990 Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res* **18**: 1415-1419.
- SWERDLOW, H., S. L. WU, H. HARKE and N. J. DOVICH, 1990 Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *J Chromatogr* **516**: 61-67.
- TARIQ, M. A., H. J. KIM, O. JEJELOWO and N. POURMAND, 2011 Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res* **39**: e120.
- TESCHENDORFF, A. E., F. MARABITA, M. LECHNER, T. BARTLETT, J. TEGNER *et al.*, 2013 A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**: 189-196.
- TESHIMA, K. M., G. COOP and M. PRZEWORSKI, 2006 How reliable are empirical genomic scans for selective sweeps? *Genome Res* **16**: 702-712.
- THORNTON, K., and P. ANDOLFATTO, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* **172**: 1607-1619.
- TRAPNELL, C., B. A. WILLIAMS, G. PERTEA, A. MORTAZAVI, G. KWAN *et al.*, 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511-515.
- VAN DIJK, E. L., Y. JASZCZYSZYN and C. THERMES, 2014 Library preparation methods for next-generation sequencing: Tone down the bias. *Exp Cell Res* **322**: 12-20.

- VOIGHT, B. F., S. KUDARAVALLI, X. WEN and J. K. PRITCHARD, 2006 A map of recent positive selection in the human genome. *PLoS Biol* **4**: e72.
- WALL, J. D., P. ANDOLFATTO and M. PRZEWORSKI, 2002 Testing models of selection and demography in *Drosophila simulans*. *Genetics* **162**: 203-216.
- WANG, J., W. WANG, R. LI, Y. LI, G. TIAN *et al.*, 2008 The diploid genome sequence of an Asian individual. *Nature* **456**: 60-65.
- WANG, Z., M. GERSTEIN and M. SNYDER, 2009 RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57-63.
- WERZNER, A., P. PAVLIDIS, L. OMETTO, W. STEPHAN and S. LAURENT, 2013 Selective sweep in the Flotillin-2 region of European *Drosophila melanogaster*. *PLoS One* **8**: e56629.
- WHEELER, D. A., M. SRINIVASAN, M. EGHOLM, Y. SHEN, L. CHEN *et al.*, 2008 The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872-876.
- WILLIAMSON, S. H., R. HERNANDEZ, A. FLEDEL-ALON, L. ZHU, R. NIELSEN *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U S A* **102**: 7882-7887.
- WILLIAMSON, S. H., M. J. HUBISZ, A. G. CLARK, B. A. PAYSEUR, C. D. BUSTAMANTE *et al.*, 2007 Localizing recent adaptive evolution in the human genome. *PLoS Genet* **3**: e90.
- WONG, W. S., and R. NIELSEN, 2004 Detecting selection in noncoding regions of nucleotide sequences. *Genetics* **167**: 949-958.
- WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI, J. F. DOEBLEY *et al.*, 2005 The effects of artificial selection on the maize genome. *Science* **308**: 1310-1314.
- XIA, Q., Y. GUO, Z. ZHANG, D. LI, Z. XUAN *et al.*, 2009 Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* **326**: 433-436.
- YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555-556.
- YANG, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586-1591.
- YE, K., M. H. SCHULZ, Q. LONG, R. APWEILER and Z. NING, 2009 Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865-2871.
- ZENG, K., S. SHI and C. I. WU, 2007 Compound tests for the detection of hitchhiking under positive selection. *Mol Biol Evol* **24**: 1898-1908.
- ZHANG, J., R. CHIODINI, A. BADR and G. ZHANG, 2011 The impact of next-generation sequencing on genomics. *J Genet Genomics* **38**: 95-109.



- AGARWAL, A., D. KOPPSTEIN, J. ROZOWSKY, A. SBONER, L. HABEGGER *et al.*, 2010 Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. *BMC Genomics* **11**: 383.
- AHN, S. M., T. H. KIM, S. LEE, D. KIM, H. GHANG *et al.*, 2009 The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19**: 1622-1629.
- AIRD, D., M. G. ROSS, W. S. CHEN, M. DANIELSSON, T. FENNELL *et al.*, 2011 Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**: R18.
- AKEY, J. M., G. ZHANG, K. ZHANG, L. JIN and M. D. SHRIVER, 2002 Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* **12**: 1805-1814.
- ANDOLFATTO, P., 2007 Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res* **17**: 1755-1762.
- ANDOLFATTO, P., and M. PRZEWORSKI, 2000 A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics* **156**: 257-268.
- AUER, P. L., and R. W. DOERGE, 2010 Statistical design and analysis of RNA sequencing data. *Genetics* **185**: 405-416.
- BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**: 519-520.
- BEGUN, D. J., and H. A. LINDFORS, 2005 Rapid evolution of genomic Acp complement in the melanogaster subgroup of *Drosophila*. *Mol Biol Evol* **22**: 2010-2021.
- BENTLEY, D. R., S. BALASUBRAMANIAN, H. P. SWERDLOW, G. P. SMITH, J. MILTON *et al.*, 2008 Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.
- BHANGALE, T. R., M. J. RIEDER and D. A. NICKERSON, 2008 Estimating coverage and power for genetic association studies using near-complete variation data. *Nat Genet* **40**: 841-843.
- BRADNAM, K. R., J. N. FASS, A. ALEXANDROV, P. BARANAY, M. BECHNER *et al.*, 2013 Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2**: 10.
- BULLARD, J. H., E. PURDOM, K. D. HANSEN and S. DUDOIT, 2010 Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**: 94.
- BUSTAMANTE, C. D., R. NIELSEN, S. A. SAWYER, K. M. OLSEN, M. D. PURUGGANAN *et al.*, 2002 The cost of inbreeding in *Arabidopsis*. *Nature* **416**: 531-534.

- CARLSON, C. S., D. J. THOMAS, M. A. EBERLE, J. E. SWANSON, R. J. LIVINGSTON *et al.*, 2005 Genomic regions exhibiting positive selection identified from dense genotype data. *Genome Res* **15**: 1553-1565.
- CHARLESWORTH, J., and A. EYRE-WALKER, 2008 The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol* **25**: 1007-1015.
- CLARK, A. G., S. GLANOWSKI, R. NIELSEN, P. D. THOMAS, A. KEJARIWAL *et al.*, 2003 Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**: 1960-1963.
- CORE, L. J., J. J. WATERFALL and J. T. LIS, 2008 Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**: 1845-1848.
- CRAWFORD, J. E., and B. P. LAZZARO, 2012 Assessing the accuracy and power of population genetic inference from low-pass next-generation sequencing data. *Front Genet* **3**: 66.
- CRISCI, J. L., Y. P. POH, S. MAHAJAN and J. D. JENSEN, 2013 The impact of equilibrium assumptions on tests of selection. *Front Genet* **4**: 235.
- DAINES, B., H. WANG, Y. LI, Y. HAN, R. GIBBS *et al.*, 2009 High-throughput multiplex sequencing to discover copy number variants in *Drosophila*. *Genetics* **182**: 935-941.
- DEPRISTO, M. A., E. BANKS, R. POPLIN, K. V. GARIMELLA, J. R. MAGUIRE *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491-498.
- DOHM, J. C., C. LOTTAZ, T. BORODINA and H. HIMMELBAUER, 2008 Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* **36**: e105.
- DONIGER, S. W., H. S. KIM, D. SWAIN, D. CORCUERA, M. WILLIAMS *et al.*, 2008 A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet* **4**: e1000183.
- DUMONT, V. B., and C. F. AQUADRO, 2005 Multiple signatures of positive selection downstream of notch on the X chromosome in *Drosophila melanogaster*. *Genetics* **171**: 639-653.
- DUMONT, V. B., J. C. FAY, P. P. CALABRESE and C. F. AQUADRO, 2004 DNA variability and divergence at the notch locus in *Drosophila melanogaster* and *D. simulans*: a case of accelerated synonymous site divergence. *Genetics* **167**: 171-185.
- EARL, D., K. BRADNAM, J. ST JOHN, A. DARLING, D. LIN *et al.*, 2011 Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* **21**: 2224-2241.
- EMDE, A. K., M. H. SCHULZ, D. WEESE, R. SUN, M. VINGRON *et al.*, 2012 Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS. *Bioinformatics* **28**: 619-627.

- ERSOZ, E. S., M. H. WRIGHT, J. L. PANGILINAN, M. J. SHEEHAN, C. TOBIAS *et al.*, 2012 SNP discovery with EST and NextGen sequencing in switchgrass (*Panicum virgatum* L.). *PLoS One* **7**: e44112.
- EWING, B., and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**: 186-194.
- EWING, B., L. HILLIER, M. C. WENDL and P. GREEN, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**: 175-185.
- FISTON-LAVIER, A. S., N. D. SINGH, M. LIPATOV and D. A. PETROV, 2010 *Drosophila melanogaster* recombination rate calculator. *Gene* **463**: 18-20.
- GALTIER, N., F. DEPAULIS and N. H. BARTON, 2000 Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics* **155**: 981-987.
- GENOMES PROJECT, C., G. R. ABECASIS, D. ALTSHULER, A. AUTON, L. D. BROOKS *et al.*, 2010 A map of human genome variation from population-scale sequencing. *Nature* **467**: 1061-1073.
- GENOMES PROJECT, C., G. R. ABECASIS, A. AUTON, L. D. BROOKS, M. A. DEPRISTO *et al.*, 2012 An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**: 56-65.
- GLINKA, S., L. OMETTO, S. MOUSSET, W. STEPHAN and D. DE LORENZO, 2003 Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* **165**: 1269-1278.
- GRAVELEY, B. R., A. N. BROOKS, J. W. CARLSON, M. O. DUFF, J. M. LANDOLIN *et al.*, 2011 The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471**: 473-479.
- GROSSMAN, S. R., I. SHLYAKHTER, E. K. KARLSSON, E. H. BYRNE, S. MORALES *et al.*, 2010 A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**: 883-886.
- GUO, Y., J. LI, C. I. LI, J. LONG, D. C. SAMUELS *et al.*, 2012 The effect of strand bias in Illumina short-read sequencing data. *BMC Genomics* **13**: 666.
- GUTTMAN, M., M. GARBER, J. Z. LEVIN, J. DONAGHEY, J. ROBINSON *et al.*, 2010 Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**: 503-510.
- HABEGGER, L., A. SBONER, T. A. GIANOULIS, J. ROZOWSKY, A. AGARWAL *et al.*, 2011 RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries. *Bioinformatics* **27**: 281-283.
- HANSEN, K. D., S. E. BRENNER and S. DUDOIT, 2010 Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**: e131.
- HELLMANN, I., Y. MANG, Z. GU, P. LI, F. M. DE LA VEGA *et al.*, 2008 Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res* **18**: 1020-1029.

- HELYAR, S. J., M. T. LIMBORG, D. BEKKEVOLD, M. BABBUCCI, J. VAN HOUDT *et al.*, 2012 SNP discovery using Next Generation Transcriptomic Sequencing in Atlantic herring (*Clupea harengus*). *PLoS One* **7**: e42089.
- HERNANDEZ, R. D., 2008 A flexible forward simulator for populations subject to selection and demography. *Bioinformatics* **24**: 2786-2787.
- HILLIER, L. W., V. REINKE, P. GREEN, M. HIRST, M. A. MARRA *et al.*, 2009 Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. *Genome Res* **19**: 657-666.
- HOLLOWAY, A. K., D. J. BEGUN, A. SIEPEL and K. S. POLLARD, 2008 Accelerated sequence divergence of conserved genomic elements in *Drosophila melanogaster*. *Genome Res* **18**: 1592-1601.
- HUDSON, R. R., 2002 Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**: 337-338.
- HUDSON, R. R., M. KREITMAN and M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153-159.
- HUELSENBECK, J. P., and F. RONQUIST, 2001 MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754-755.
- HUNKAPILLER, T., R. J. KAISER, B. F. KOOP and L. HOOD, 1991 Large-scale and automated DNA sequence determination. *Science* **254**: 59-67.
- HUTCHISON, C. A., 3RD, 2007 DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res* **35**: 6227-6237.
- IRVIN, S. D., K. A. WETTERSTRAND, C. M. HUTTER and C. F. AQUADRO, 1998 Genetic variation and differentiation at microsatellite loci in *Drosophila simulans*. Evidence for founder effects in new world populations. *Genetics* **150**: 777-790.
- JENSEN, J. D., V. L. BAUER DUMONT, A. B. ASHMORE, A. GUTIERREZ and C. F. AQUADRO, 2007a Patterns of sequence variability and divergence at the diminutive gene region of *Drosophila melanogaster*: complex patterns suggest an ancestral selective sweep. *Genetics* **177**: 1071-1085.
- JENSEN, J. D., Y. KIM, V. B. DUMONT, C. F. AQUADRO and C. D. BUSTAMANTE, 2005 Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**: 1401-1410.
- JENSEN, J. D., K. R. THORNTON and P. ANDOLFATTO, 2008a An approximate bayesian estimator suggests strong, recurrent selective sweeps in *Drosophila*. *PLoS Genet* **4**: e1000198.
- JENSEN, J. D., K. R. THORNTON and C. F. AQUADRO, 2008b Inferring selection in partially sequenced regions. *Mol Biol Evol* **25**: 438-446.
- JENSEN, J. D., K. R. THORNTON, C. D. BUSTAMANTE and C. F. AQUADRO, 2007b On the utility of linkage disequilibrium as a statistic for identifying targets of positive selection in nonequilibrium populations. *Genetics* **176**: 2371-2379.

- JIANG, R., S. TAVARE and P. MARJORAM, 2009 Population genetic inference from resequencing data. *Genetics* **181**: 187-197.
- JOHNSON, P. L., and M. SLATKIN, 2006 Inference of population genetic parameters in metagenomics: a clean look at messy data. *Genome Res* **16**: 1320-1327.
- JOHNSON, P. L., and M. SLATKIN, 2008 Accounting for bias from sequencing error in population genetic estimates. *Mol Biol Evol* **25**: 199-206.
- JORDAN, K. W., M. A. CARBONE, A. YAMAMOTO, T. J. MORGAN and T. F. MACKAY, 2007 Quantitative genomics of locomotor behavior in *Drosophila melanogaster*. *Genome Biol* **8**: R172.
- JUAREZ, M. T., R. A. PATTERSON, E. SANDOVAL-GUILLEN and W. MCGINNIS, 2011 Duox, Flotillin-2, and Src42A are required to activate or delimit the spread of the transcriptional response to epidermal wounds in *Drosophila*. *PLoS Genet* **7**: e1002424.
- KAMPA, D., J. CHENG, P. KAPRANOV, M. YAMANAKA, S. BRUBAKER *et al.*, 2004 Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* **14**: 331-342.
- KATANAEV, V. L., G. P. SOLIS, G. HAUSMANN, S. BUESTORF, N. KATANAYEVA *et al.*, 2008 Reggie-1/flotillin-2 promotes secretion of the long-range signalling forms of Wingless and Hedgehog in *Drosophila*. *EMBO J* **27**: 509-521.
- KEIGHTLEY, P. D., and A. EYRE-WALKER, 2007 Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177**: 2251-2261.
- KEIGHTLEY, P. D., U. TRIVEDI, M. THOMSON, F. OLIVER, S. KUMAR *et al.*, 2009 Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. *Genome Res* **19**: 1195-1201.
- KELLEY, J. L., and W. J. SWANSON, 2008 Positive selection in the human genome: from genome scans to biological significance. *Annu Rev Genomics Hum Genet* **9**: 143-160.
- KIM, S. Y., K. E. LOHMEYER, A. ALBRECHTSEN, Y. LI, T. KORNELIUSSEN *et al.*, 2011 Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* **12**: 231.
- KIM, Y., and R. NIELSEN, 2004 Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**: 1513-1524.
- KIM, Y., and W. STEPHAN, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**: 765-777.
- KOBOLDT, D. C., K. CHEN, T. WYLIE, D. E. LARSON, M. D. MCLELLAN *et al.*, 2009 VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**: 2283-2285.

- KOBOLDT, D. C., K. M. STEINBERG, D. E. LARSON, R. K. WILSON and E. R. MARDIS, 2013 The next-generation sequencing revolution and its impact on genomics. *Cell* **155**: 27-38.
- KORFHAGE, C., E. FISCH, E. FRICKE, S. BAEDKER and D. LOEFFERT, 2013 Whole-genome amplification of single-cell genomes for next-generation sequencing. *Curr Protoc Mol Biol* **104**: 7 14 11-17 14 11.
- KREITMAN, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**: 412-417.
- LANGLEY, C. H., K. STEVENS, C. CARDENO, Y. C. LEE, D. R. SCHRIDER *et al.*, 2012 Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* **192**: 533-598.
- LANGMEAD, B., C. TRAPNELL, M. POP and S. L. SALZBERG, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.
- LEPROUST, E., 2012 Target enrichment strategies for next generation sequencing. *MLO Med Lab Obs* **44**: 26-27.
- LEVIN, J. Z., M. YASSOUR, X. ADICONIS, C. NUSBAUM, D. A. THOMPSON *et al.*, 2010 Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7**: 709-715.
- LI, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987-2993.
- LI, H., and R. DURBIN, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754-1760.
- LI, H., B. HANDSAKER, A. WYSOKER, T. FENNEL, J. RUAN *et al.*, 2009a The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.
- LI, H., J. RUAN and R. DURBIN, 2008a Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**: 1851-1858.
- LI, H., and W. STEPHAN, 2005 Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome. *Genetics* **171**: 377-384.
- LI, H., and W. STEPHAN, 2006 Inferring the demographic history and rate of adaptive substitution in *Drosophila*. *PLoS Genet* **2**: e166.
- LI, J., H. JIANG and W. H. WONG, 2010a Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biol* **11**: R50.
- LI, R., Y. LI, K. KRISTIANSEN and J. WANG, 2008b SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**: 713-714.
- LI, R., C. YU, Y. LI, T. W. LAM, S. M. YIU *et al.*, 2009b SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**: 1966-1967.
- LI, Y., S. L. YANG, Z. L. TANG, W. T. CUI, Y. L. MU *et al.*, 2010b Expression and SNP association analysis of porcine FBXL4 gene. *Mol Biol Rep* **37**: 579-585.

- LIGHTEN, J., C. VAN OOSTERHOUT, I. G. PATERSON, M. McMULLAN and P. BENTZEN, 2014 Ultra-deep Illumina sequencing accurately identifies MHC class IIb alleles and provides evidence for copy number variation in the guppy (*Poecilia reticulata*). *Mol Ecol Resour.*
- LIU, C. M., T. WONG, E. WU, R. LUO, S. M. YIU *et al.*, 2012 SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics* **28**: 878-879.
- LYNCH, M., 2008 Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol Biol Evol* **25**: 2409-2419.
- LYNCH, M., 2009 Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* **182**: 295-301.
- MACKAY, T. F., S. RICHARDS, E. A. STONE, A. BARBADILLA, J. F. AYROLES *et al.*, 2012 The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**: 173-178.
- MAMANOVA, L., A. J. COFFEY, C. E. SCOTT, I. KOZAREWA, E. H. TURNER *et al.*, 2010 Target-enrichment strategies for next-generation sequencing. *Nat Methods* **7**: 111-118.
- MARCHINI, J., and B. HOWIE, 2010 Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**: 499-511.
- MARDIS, E. R., 2008 Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9**: 387-402.
- MARGUERAT, S., and J. BAHLER, 2010 RNA-seq: from technology to biology. *Cell Mol Life Sci* **67**: 569-579.
- MARGULIES, M., M. EGHOLM, W. E. ALTMAN, S. ATTIYA, J. S. BADER *et al.*, 2005 Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.
- MARIONI, J. C., M. WHITE, S. TAVARE and A. G. LYNCH, 2008 Hidden copy number variation in the HapMap population. *Proc Natl Acad Sci U S A* **105**: 10067-10072.
- MARYGOLD, S. J., P. C. LEYLAND, R. L. SEAL, J. L. GOODMAN, J. THURMOND *et al.*, 2013 FlyBase: improvements to the bibliography. *Nucleic Acids Res* **41**: D751-757.
- McKENNA, A., M. HANNA, E. BANKS, A. SIVACHENKO, K. CIBULSKIS *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.
- MORTAZAVI, A., B. A. WILLIAMS, K. MCCUE, L. SCHAEFFER and B. WOLD, 2008 Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**: 621-628.
- NAGALAKSHMI, U., K. WAERN and M. SNYDER, 2010 RNA-Seq: a method for comprehensive transcriptome analysis. *Curr Protoc Mol Biol* **Chapter 4**: Unit 4 11 11-13.
- NECHAEV, S., D. C. FARGO, G. DOS SANTOS, L. LIU, Y. GAO *et al.*, 2010 Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327**: 335-338.

- NI, T., D. L. CORCORAN, E. A. RACH, S. SONG, E. P. SPANA *et al.*, 2010 A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* **7**: 521-527.
- NIELSEN, R., C. BUSTAMANTE, A. G. CLARK, S. GLANOWSKI, T. B. SACKTON *et al.*, 2005a A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.
- NIELSEN, R., I. HELLMANN, M. HUBISZ, C. BUSTAMANTE and A. G. CLARK, 2007 Recent and ongoing selection in the human genome. *Nat Rev Genet* **8**: 857-868.
- NIELSEN, R., M. J. HUBISZ, I. HELLMANN, D. TORGERSON, A. M. ANDRES *et al.*, 2009 Darwinian and demographic forces affecting human protein coding genes. *Genome Res* **19**: 838-849.
- NIELSEN, R., S. WILLIAMSON, Y. KIM, M. J. HUBISZ, A. G. CLARK *et al.*, 2005b Genomic scans for selective sweeps using SNP data. *Genome Res* **15**: 1566-1575.
- OLEKSYK, T. K., K. ZHAO, F. M. DE LA VEGA, D. A. GILBERT, S. J. O'BRIEN *et al.*, 2008 Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS One* **3**: e1712.
- OMETTO, L., S. GLINKA, D. DE LORENZO and W. STEPHAN, 2005 Inferring the effects of demography and selection on *Drosophila melanogaster* populations from a chromosome-wide scan of DNA variation. *Mol Biol Evol* **22**: 2119-2130.
- ONMUS-LEONE, F., J. HANG, R. J. CLIFFORD, Y. YANG, M. C. RILEY *et al.*, 2013 Enhanced de novo assembly of high throughput pyrosequencing data using whole genome mapping. *PLoS One* **8**: e61762.
- OSSOWSKI, S., K. SCHNEEBERGER, R. M. CLARK, C. LANZ, N. WARTHMAN *et al.*, 2008 Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* **18**: 2024-2033.
- OZSOLAK, F., and P. M. MILOS, 2011 RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12**: 87-98.
- PAREEK, C. S., R. SMOCZYNSKI and A. TRETYN, 2011 Sequencing technologies and genome sequencing. *J Appl Genet* **52**: 413-435.
- PARK, P. J., 2009 ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**: 669-680.
- PAVLIDIS, P., J. D. JENSEN and W. STEPHAN, 2010 Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics* **185**: 907-922.
- PAVLIDIS, P., D. ZIVKOVIC, A. STAMATAKIS and N. ALACHIOTIS, 2013 SweeD: likelihood-based detection of selective sweeps in thousands of genomes. *Mol Biol Evol* **30**: 2224-2234.
- PEPKE, S., B. WOLD and A. MORTAZAVI, 2009 Computation for ChIP-seq and RNA-seq studies. *Nat Methods* **6**: S22-32.



- POOL, J. E., and C. F. AQUADRO, 2006 History and structure of sub-Saharan populations of *Drosophila melanogaster*. *Genetics* **174**: 915-929.
- POOL, J. E., R. B. CORBETT-DETIG, R. P. SUGINO, K. A. STEVENS, C. M. CARDENO *et al.*, 2012 Population Genomics of sub-saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genet* **8**: e1003080.
- POOL, J. E., I. HELLMANN, J. D. JENSEN and R. NIELSEN, 2010 Population genetic inference from genomic sequence variation. *Genome Res* **20**: 291-300.
- PORCU, E., S. SANNA, C. FUCHSBERGER and L. G. FRITSCH, 2013 Genotype imputation in genome-wide association studies. *Curr Protoc Hum Genet* **Chapter 1**: Unit 1 25.
- PRESGRAVES, D. C., 2006 Intron length evolution in *Drosophila*. *Mol Biol Evol* **23**: 2203-2213.
- RIVERA, C. M., and B. REN, 2013 Mapping human epigenomes. *Cell* **155**: 39-55.
- ROBINSON, M. D., D. J. MCCARTHY and G. K. SMYTH, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139-140.
- ROYCE, T. E., J. S. ROZOWSKY, P. BERTONE, M. SAMANTA, V. STOLC *et al.*, 2005 Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet* **21**: 466-475.
- SABETI, P. C., D. E. REICH, J. M. HIGGINS, H. Z. LEVINE, D. J. RICHTER *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**: 832-837.
- SABETI, P. C., S. F. SCHAFFNER, B. FRY, J. LOHMUELLER, P. VARILLY *et al.*, 2006 Positive natural selection in the human lineage. *Science* **312**: 1614-1620.
- SABETI, P. C., P. VARILLY, B. FRY, J. LOHMUELLER, E. HOSTETTER *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913-918.
- SANGER, F., 1988 Sequences, sequences, and sequences. *Annu Rev Biochem* **57**: 1-28.
- SCHWARTZ, T. S., H. TAE, Y. YANG, K. MOCKAITIS, J. L. VAN HEMERT *et al.*, 2010 A garter snake transcriptome: pyrosequencing, de novo assembly, and sex-specific differences. *BMC Genomics* **11**: 694.
- SHAPIRO, E., T. BIEZUNER and S. LINNARSSON, 2013 Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat Rev Genet* **14**: 618-630.
- SHARON, D., H. TILGNER, F. GRUBERT and M. SNYDER, 2013 A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**: 1009-1014.
- SHEN, Y., Z. WAN, C. COARFA, R. DRABEK, L. CHEN *et al.*, 2010 A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res* **20**: 273-280.
- SHENDURE, J., R. D. MITRA, C. VARMA and G. M. CHURCH, 2004 Advanced sequencing technologies: methods and goals. *Nat Rev Genet* **5**: 335-344.

- SINGH, N. D., P. F. ARNDT and D. A. PETROV, 2005 Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. *Genetics* **169**: 709-722.
- SINGH, N. D., J. D. JENSEN, A. G. CLARK and C. F. AQUADRO, 2013 Inferences of demography and selection in an African population of *Drosophila melanogaster*. *Genetics* **193**: 215-228.
- SMITH, J. M., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. *Genet Res* **23**: 23-35.
- STAPLEY, J., J. REGER, P. G. FEULNER, C. SMADJA, J. GALINDO *et al.*, 2010 Adaptation genomics: the next generation. *Trends Ecol Evol* **25**: 705-712.
- STEPHAN, W., and H. LI, 2007 The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity (Edinb)* **98**: 65-68.
- STEPHAN, W., Y. S. SONG and C. H. LANGLEY, 2006 The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics* **172**: 2647-2663.
- STONE, E. A., 2012 Joint genotyping on the fly: identifying variation among a sequenced panel of inbred lines. *Genome Res* **22**: 966-974.
- SWERDLOW, H., and R. GESTELAND, 1990 Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res* **18**: 1415-1419.
- SWERDLOW, H., S. L. WU, H. HARKE and N. J. DOVICH, 1990 Capillary gel electrophoresis for DNA sequencing. Laser-induced fluorescence detection with the sheath flow cuvette. *J Chromatogr* **516**: 61-67.
- TARIQ, M. A., H. J. KIM, O. JEJELOWO and N. POURMAND, 2011 Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res* **39**: e120.
- TESCHENDORFF, A. E., F. MARABITA, M. LECHNER, T. BARTLETT, J. TEGNER *et al.*, 2013 A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**: 189-196.
- TESHIMA, K. M., G. COOP and M. PRZEWORSKI, 2006 How reliable are empirical genomic scans for selective sweeps? *Genome Res* **16**: 702-712.
- THORNTON, K., and P. ANDOLFATTO, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. *Genetics* **172**: 1607-1619.
- TRAPNELL, C., B. A. WILLIAMS, G. PERTEA, A. MORTAZAVI, G. KWAN *et al.*, 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**: 511-515.
- VAN DIJK, E. L., Y. JASZCZYSZYN and C. THERMES, 2014 Library preparation methods for next-generation sequencing: Tone down the bias. *Exp Cell Res* **322**: 12-20.
- VOIGHT, B. F., S. KUDARAVALLI, X. WEN and J. K. PRITCHARD, 2006 A map of recent positive selection in the human genome. *PLoS Biol* **4**: e72.
- WALL, J. D., P. ANDOLFATTO and M. PRZEWORSKI, 2002 Testing models of selection and demography in *Drosophila simulans*. *Genetics* **162**: 203-216.

- WANG, J., W. WANG, R. LI, Y. LI, G. TIAN *et al.*, 2008 The diploid genome sequence of an Asian individual. *Nature* **456**: 60-65.
- WANG, Z., M. GERSTEIN and M. SNYDER, 2009 RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**: 57-63.
- WERZNER, A., P. PAVLIDIS, L. OMETTO, W. STEPHAN and S. LAURENT, 2013 Selective sweep in the Flotillin-2 region of European *Drosophila melanogaster*. *PLoS One* **8**: e56629.
- WHEELER, D. A., M. SRINIVASAN, M. EGHOLM, Y. SHEN, L. CHEN *et al.*, 2008 The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872-876.
- WILLIAMSON, S. H., R. HERNANDEZ, A. FLEDEL-ALON, L. ZHU, R. NIELSEN *et al.*, 2005 Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proc Natl Acad Sci U S A* **102**: 7882-7887.
- WILLIAMSON, S. H., M. J. HUBISZ, A. G. CLARK, B. A. PAYSEUR, C. D. BUSTAMANTE *et al.*, 2007 Localizing recent adaptive evolution in the human genome. *PLoS Genet* **3**: e90.
- WONG, W. S., and R. NIELSEN, 2004 Detecting selection in noncoding regions of nucleotide sequences. *Genetics* **167**: 949-958.
- WRIGHT, S. I., I. V. BI, S. G. SCHROEDER, M. YAMASAKI, J. F. DOEBLEY *et al.*, 2005 The effects of artificial selection on the maize genome. *Science* **308**: 1310-1314.
- XIA, Q., Y. GUO, Z. ZHANG, D. LI, Z. XUAN *et al.*, 2009 Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). *Science* **326**: 433-436.
- YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555-556.
- YANG, Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586-1591.
- YE, K., M. H. SCHULZ, Q. LONG, R. APWEILER and Z. NING, 2009 Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**: 2865-2871.
- ZENG, K., S. SHI and C. I. WU, 2007 Compound tests for the detection of hitchhiking under positive selection. *Mol Biol Evol* **24**: 1898-1908.
- ZHANG, J., R. CHIODINI, A. BADR and G. ZHANG, 2011 The impact of next-generation sequencing on genomics. *J Genet Genomics* **38**: 95-109.